Opinnäytetyö (AMK)

Liiketalouden koulutusohjelma

2019

Olga Alto

BIG DATAN KÄYTTÖ LIIKETOIMINNAN ENNUSTAMISEEN: TIELIIKENNEONNETTOMUUDET SUOMESSA

TURKU AMK
TURKU UNIVERSITY OF APPLIED SCIENCES

OPINNÄYTETYÖ (AMK) | TIIVISTELMÄ

TURUN AMMATTIKORKEAKOULU

Liiketalouden koulutusohjelma

2019 | 36 sivua

Olga Alto

BIG DATAN KÄYTTÖ LIIKETOIMINNAN ENNUSTAMISEEN: TIELIIKENNEONNETTOMUUDET SUOMESSA

Tämän opinnäytetyön tarkoituksena on selvittää, mitä tietoja voidaan ennustaa suurista tietomääristä. Aineistona on käytetty Suomessa liikennetapaturmia koskevia avoimia lähteitä vuosilta 2015 – 2017.

Työssä ennustetaan liikenneonnettomuuksien tulevia arvoja vuonna 2018. Regressioanalyysin avulla arvioitiin kuljettajan iän, ajokokemuksen ja nopeusrajoitusten vaikutusta loukkaantuneiden lukumäärään liikenneonnettomuuksissa. Luottamusväli luotiin, jotta voitaisiin arvioida uhrien lukumäärä tieliikenneonnettomuuksissa kuljettajan iän ja ajokokemuksen mukaan.

Tutkimuksessa ennustettiin, että tieliikenneonnettomuuksien määrä vuonna 2018 pysyi vuoden 2017 tasolla, mutta vähemmän kuin vuosina 2015-2016.

Regressioanalyysin avulla todettiin, että ajokokemuksella on vahva vaikutus uhrien määrään. Liikenneonnettomuudet vaikuttavat paljon nuorempiin kuljettajiin, joilla on vähän ajokokemusta. Ikääntyneihin kuljettajiin tämä vaikutus vähenee. Nopeusrajoitusten ja uhrien lukumäärän suhdetta regressiomalleja verrattaessa pidettiin heikkona.

Regressioanalyysin aikana luotiin luottamusväliä ennustetulle uhrien määrälle 19-vuotiaalla kuljettajalla, jolla oli yhden vuoden ajokokemus, 30-vuotias kuljettaja, jolla on yhdeksän vuoden kokemus, 50-vuotias kuljettaja, jolla on 19 vuoden kokemus. Ennusteiden mukaan luottamusväli laskee (uhrien määrä pienenee), kun ajokokemus ja kuljettajan ikä kasvaavat. Niinpä, käyttämällä regressioanalyysiä ja rakentamalla luottamusvälejä ennustetun uhrien lukumäärän kanssa, voidaan päätellä, että ajokokemus ja kuljettajan ikä vaikuttavat suuresti uhrien määrään.

Tutkimuksen tulos voi olla hyödyllistä vakuutusyhtiölle vakuutusriskien arvioinnissa ja vakuutuskustannusten laskemisessa.

Tutkimus suoritettiin R-ohjelmointikielellä RStudio-ympäristössä. Työssä käytettiin tilastotietoja, ohjelmointia, tiedon louhintamenetelmiä.

ASIASANAT:

massadata, avoin data, data-analyysi, tiedonlouhinta, R-ohjelmointikieli.

BACHELOR´S THESIS | ABSTRACT

TURKU UNIVERSITY OF APPLIED SCIENCES

Business Administration

2019 | 36 pages

Olga Alto

USING BIG DATA FOR PREDICTION IN BUSINESS: ROAD TRAFFIC ACCIDENTS IN FINLAND

The purpose of this thesis is to find out what information can be predicted from large amounts of data, for which open data on road traffic accidents in Finland in 2015 – 2017 has been used.

The study predicted the future values of traffic accidents in 2018. Using the regression analysis, the influence of the driver's age, driving experience and speed limits on the number of people injured in road traffic accidents is estimated. 95 % confidence intervals were built to estimate the number of victims, depending on the set age and driving experience.

The study predicted that the number of road accidents in 2018 remained at the 2017 level, but less than in 2015-2016.

Regression analysis found that driving experience has a strong impact on the number of victims. Traffic accidents affect much younger drivers with little driving experience. For older drivers, this effect is reduced. The relationship between speed limits and the number of victims compared to regression models was considered weak.

During the regression analysis, a confidence interval was created for the predicted number of victims with a 19-year-old driver with a one-year driving experience, a 30-year-old driver with nine years experience, a 50-year-old driver with 19 years of experience. According to forecasts, the confidence interval will decrease (the number of victims will decrease) as the driving experience and driver age increase. Thus, by using regression analysis and building confidence intervals with the predicted number of victims, it can be concluded that the driving experience and driver age have a major impact on the number of victims.

The result of the study may be useful for the insurance company in assessing insurance risks and calculating insurance costs.

The research was conducted in the R programming language in RStudio and used information about statistics, programming, date mining methods.

KEYWORDS:

Big data, open data, data analysis, data mining, R programming language.

SISÄLTÖ

KUVAT

1 JOHDANTO

R-ohjelmointikielen tietojen analysoinnin yhteydessä havahduttiin Big datan käytön laajentamiseen. Kysymykseksi nousi, mitä tietoja voidaan ennustaa käyttämällä tiedon louhinnan menetelmiä. Tähän opinnäytetyöhön valittiin tieliikenneonnettomuustietokanta Suomessa vuosilta 2015–2017.

Tämän opinnäytetyössä avataan Big datan käsitettä ja rakennetta. Tavoitteena on analysoida massadatan rooli liiketoiminnassa ja esittää esikäsittelytietojen vaiheet RStudiossa. Työssä käytetään regressioanalyysiä kuljettajan iän ja ajokokemuksen vaikutuksen arvioimiseksi liikenneonnettomuuksissa loukkaantuneiden ihmisten määrään. Toisessa tarkastelussa lisäparametrinä on nopeusrajoitus, jolloin regressioanalyysiä käytetään kuljettajan iän, ajokokemuksen ja nopeusrajoituksen vaikutuksen arvioimiseksi liikenneonnettomuuksissa loukkaantuneiden ihmisten määrään.

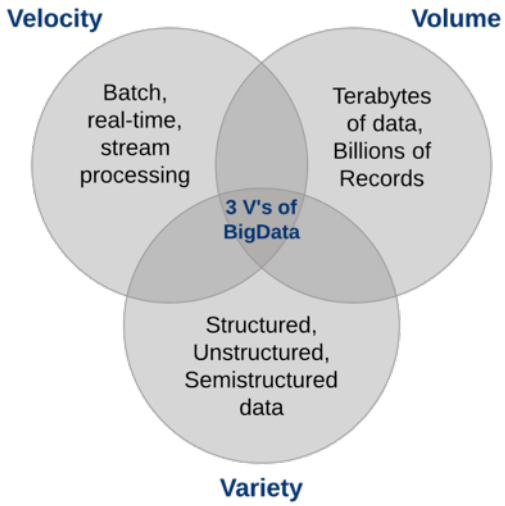
Työssä rakennetaan tilastollisen päättelyn luottamusvälejä uhrien määrän ennakoimiseksi liikenneonnettomuuksissa. Tavoitteena on rakentaa kolme mallia ennustamaan liikenneonnettomuuksien määrän tulevia arvoja, testata niitä ja valita paras.

Koska opinnäytetyön aikana kirjoitettaessa ei julkaistu tietoja liikenneonnettomuuksista vuonna 2018, ennustejaksoksi valittiin 1.2018–12.2018. Työssä analysoidaan mahdollista käytännön hyötyä vakuutuslaitoksille.

2 MITÄ ON BIG DATA

2.1 Big datan käsite

Big datan käsitys tuli tutuksi jo vuoden 2015 tietämillä, mutta nykyinen big data -hype lähti nousuun vuonna 2011. Tarkkaa alkamisajankohtaa sille, koska big datasta alettiin puhua, ei ole mahdollista määritellä. Vuonna 2001 META Group -yrityksen työntekijä Goug Laney julkaisi muutaman sivun mittaisen raportin, jossa puhutaan datamäärien ja niiden sisällön vaihtelevuuden kasvusta tulevaisuudessa. Tässä yhteydessä mainittiin kolme V-kirjaimella alkanutta sanaa volume, variety ja velocity, jotka kääntyvät suomeksi myös V-kirjaimilla alkaviksi: volyymi, vaihtelevuus ja vauhti. (Salo 2014.)



Kuva 1. Big datan 3V:tä.

Volyymilla viitataan havaittuun ongelmaan, että datan määrä maailmassa kasvaa eksponentiaalisesti. Vauhdilla puolestaan tarkoitetaan kiihtyvää nopeutta, jolla dataa syötetään tietojärjestelmiin ja jolla sitä täytyisi sieltä myös saada käyttöön. Vaihtelevuus kuvaa datan muuttumista yhä heterogeenisemmaksi sen lähteiden monipuolistuessa. (Salo 2013.)

2.2 Big datan rakenne

Ihminen pyrkii luokittelemaan asioita. Big Data on laajasti jaettu kolmeen päätyyppiin, jotka ovat

- strukturoidut tiedot (Structured data)
- puolijohdetut tiedot (Semi-structured data)
- strukturoimattomat tiedot (Unstructured data).

Strukturoituja tietoja käytetään viittaamaan tietoihin, jotka on jo tallennettu tietokantoihin, järjestyksessä. Sen osuus on noin 20 % kaikista olemassa olevista tiedoista, ja sitä käytetään eniten ohjelmoinnissa ja tietokoneisiin liittyvässä toiminnassa.

On olemassa kaksi strukturoitujen tietovälineiden ja ihmisten lähteitä. Kaikki antureilta, web-lokista ja rahoitusjärjestelmistä saadut tiedot luokitellaan konekohtaisiin tietoihin. Näitä ovat lääketieteelliset laitteet, GPS-tiedot, palvelimien ja sovellusten käyttötilastojen tiedot ja valtava määrä tietoja, jotka yleensä liikkuvat kaupankäyntialustojen kautta. Ihmisen synnyttämät strukturoidut tiedot sisältävät pääasiassa kaikki tiedot, joita ihminen syöttää tietokoneeseen, kuten hänen nimensä ja muut henkilötiedot. Kun henkilö napsauttaa linkkiä Internetissä tai jopa siirtyy peliin, tiedot luodaan - yritykset voivat käyttää tietoja selvittääkseen asiakkaiden käyttäytymista ja tehdä tarvittavat päätökset ja muutokset. (Types of Big Data 2016.)

Puolijohdetut tiedot (semi-structured data) on löyhästi määritelty rakenne. Esimerkiksi internetsivuston keräämät lokitiedot ovat tällaista. Data sisältää tietoa, joka on merkitty ennaltamäärätyllä tavalla, mutta minkä tahansa yksittäisen tiedon etsiminen saattaa vaatia mittavaa etsimistä. Analysointia varten dataa joudutaan luultavasti merkittävästi muokkaamaan ja sieltä poimimaan ne osat, jotka ovat varsinaisesti hyödyllisiä annetun kysymyksen näkökulmasta. (Semi-structured data 2019.)

Strukturoimattomat tiedot (unstructured data) ovat pohjimmiltaan tietoja, joilla ei ole ennalta määriteltyä datamallia ja / tai ei sovi hyvin relaatiotietokantaan. Strukturoimaton tieto on tyypillisesti tekstiä raskasta, mutta se voi sisältää myös tietoja, kuten päivämääriä, numeroita ja faktoja. (Minelli & Chambers 2012, 39.)

2.3 Big datan rooli liiketoiminnassa

Big datan merkitys ei ole se, kuinka paljon tietoa yrityksellä on, vaan miten yritys käyttää kerättyjä tietoja. Jokainen yritys käyttää tietoja omalla tavallaan: mitä tehokkaammin yritys käyttää tietojaan, sitä enemmän sillä on potentiaalia kasvaa. Yhtiö voi ottaa tietoja mistä tahansa lähteestä ja analysoida tietoa löytääkseen vastauksia, joiden avulla saavutetaan seuraavat tavoitteet:

- Kustannussäästöt. Jotkin Big data -työkalut, kuten Hadoop ja Cloud-Based Analytics, voivat tuoda yrityksille kustannusetuja, kun suuria määriä tietoja tallennetaan, ja nämä työkalut auttavat myös tunnistamaan tehokkaammat liiketoimintatavat.
- Ajan vähentäminen. Hadoopin ja muistin analytiikan kaltaisten työkalujen suurin nopeus voi helposti tunnistaa uusia tietolähteitä, jotka auttavat yrityksiä analysoimaan tietoja välittömästi ja tekemään nopeita päätöksiä oppimisen perusteella.
- Uusi tuotekehitys. Tutustumalla asiakkaiden tarpeiden ja tyytyväisyyden suuntauksiin analytiikan avulla voit luoda tuotteita asiakkaiden toiveiden mukaan.
- Markkinatilanteiden ymmärtäminen: Analysoimalla suuria tietoja voi ymmärtää paremmin nykyiset markkinaolosuhteet. Esimerkiksi analysoimalla asiakkaiden ostokäyttäytymistä yritys voi selvittää, mitä tuotteita myydään eniten ja jotka tuottavat tuotteita tämän suuntauksen mukaisesti. Tällä tavoin se voi olla kilpailijoidensa edessä.
- Verkkomaineen hallinta. Big-tietovälineet voivat tehdä tunnelman analyysin. Siksi voidaan saada palautetta siitä, kuka sanoo, mitä yrityksesi on. Jos haluat seurata ja parantaa yrityksesi läsnäoloa verkossa, suuret tietovälineet voivat auttaa tässä kaikissa. (5 Benefits: competitive advantages of Big Data in business 2017).

Jokainen teollisuusala oppii hyödyntämään Big Data -analyysin etuja ja näyttää varmalta, että innovatiivisten menetelmien löytäminen tietojen keräämiseksi, tallentamiseksi ja analysoimiseksi maksaa suurta osaa liiketoiminnasta lähitulevaisuudessa. (Marr 2015.)

2.4 Big data ja vakuutusyhtiöt

Big data ansaitsee erillisen huomion vakuutustoiminnassa, koska liikenneonnettomuuksien analyysi Suomessa antaa käsityksen onnettomuuksien

määrään vaikuttavista tekijöistä. Tämän perusteella voidaan kehittää vakuutusmaksukertoimia kuljettajan iästä ja ajokokemuksesta riippuen. Lisäksi liikenneonnettomuuksien määrän ja vastaavasti vakuutusmaksujen määrän ennustaminen voi auttaa ennakoimaan vakuutusorganisaation voiton kasvua (tai laskua).
Seuraavaksi analysoidaan Big datan hyötyä vakuutustoiminnassa.

Vakuutusala perustuu pääosin ennusteisiin, ja sille on ominaista korkean riskin aktiivisuusprofiili. Teollisuuden erittäin kilpailukykyinen luonne ja jatkuvasti muuttuva kuluttajakäyttäytyminen pakottavat niin pienet kuin suuretkin vakuutusyhtiöt investoimaan tarkkoihin ja tehokkaisiin tapoihin ennustaa kuluttajien käyttäytymistä ja minimoimaan näin riskit. Analytics ja suuret tietomäärät antavat vakuutusyhtiöille mahdollisuuden arvioida ja minimoida riskinsä, kun niitä käytetään oikein, mutta myös räätälöimään enemmän kuluttajien tarpeita ja odotuksia vastaavia vakuutustuotteita.

Riskinarviointi ja suuret tiedot: vakuutusyhtiöt voivat nykyään käyttää ennustavia mittauksia ja käyttää niitä arvioidakseen yrityksen riskejä ja haavoittuvuuksia. On vaikuttavaa, miten autovakuutusyhtiöt voivat käyttää auto- tai älykellonsa sosio-maantieteellisiä koordinaatteja ja ennustaa, onko vakuutuksenottaja todennäköisesti riskialtis auto-onnettomuuteen, varastetaanko hänen ajoneuvonsa tai joutuuko muutoin tapahtumien kohteeksi. Lisäksi prosessissa käytettävät tiedot ovat peräisin ajoneuvon tietoliikennelaitteista, terveysseurantajista ja ulkoisista tiedoista, kuten tien olosuhteista tai naapuriston turvallisuustasoista. Tämä on yksi vaikuttavimmista tavoista, joilla vakuutusyhtiöt voivat maksimoida tiedot, jotka yleensä tallennetaan eri välineisiin, vakuutusmaksujen räätälöimiseksi ja suunnittelemiseksi. (Mallon 2018.)

Petosten havaitseminen: petos on kasvaava ongelma vakuutussektorilla. Se maksaa teollisuudelle miljardeja dollareita. Yhdysvalloissa petokset varastavat 80 miljardia dollaria vuodessa kaikissa vakuutuslinjoissa. Hyvä uutinen on, että koneen oppimisalgoritmit voivat helposti poistaa inhimilliset virheet ja huomaamatta jääneet vilpilliset mallit tunnistamalla poikkeuksia. Ja ne voivat myös auttaa vakuutusyhtiöitä ilmoittamaan epäilyttävistä vaateista, jotka edellyttävät syvempää tutkimusta.

Vakuutusyhtiöt käyttävät yleensä ennakoivia malleja, joissa käytetään aiempia petollisia toimia. Jos vaateiden muuttujat vastaavat aiempia petostapauksia, nämä väitteet on liitetty lisätutkimuksiin. Lisäksi voidaan sisällyttää monimutkaisia mittareita, kuten väitteen tekevän henkilön ja kumppaniorganisaatioiden hienovaraisia

käyttäytymismalleja. Se lisäisi entisestään tilastollisten mallien kykyä tunnistaa petosjärjestelmät. (Ada 2018.)

Asiakkaiden segmentointi ja tarkkuusmarkkinointi: vakuutusyhtiöiden on kerättävä sekä vakuutusasiakkaita että vakuutusjärjestelmän ulkopuolisia tietoja, esimerkiksi vakuutusyhtiöitä. Kumppaneilta saadut tiedot sekä Internetistä haettavat sosiaaliset ja käyttäytymiseen liittyvät tiedot ovat tärkeitä vakuutusmarkkinoiden segmentoitumiselle. Asiakasinformaation ja käyttäytymisen perusteellisten analyysien avulla vakuutusyhtiöt ymmärtävät asiakkaiden tarpeet, jolloin vakuutuksenantajat voivat tunnistaa potentiaaliset asiakkaat ja suositella sopivia tuotteita näkymiin ja toteuttaa lopulta tarkan markkinoinnin, joka erottaa ne kilpailijoista.

Suurten tietojen aikakaudella markkinoijien tulisi harjoittaa mainostarkoituksia, jotka on räätälöity hienostuneisiin asiakassegmentteihin sen sijaan, että yrittäisivät houkutella eri kuluttajaryhmiä samaan mainokseen tai soveltaa samaa markkinointitekniikkaa. Käyttäytymisen ja ajallisen tiedon luokittelu antaa vakuutusyhtiöille mahdollisuuden ymmärtää, kuinka paljon tietty käyttäjäryhmä on kiinnostunut tietyistä tuotteista, ja siten vakuuttaa vakuutusyhtiöitä sopivimpien markkinointitekniikoiden valinnasta. Esimerkiksi terveys ja liikkuvuuteen liittyvä tapaturmavakuutus olisi saatettava markkinoille kuluttajille, jotka viettävät yli viisi tuntia vuorokaudessa langattomassa internetissä. Kiittisen sairausvakuutuksen tulisi olla suositeltavaa lihansyöjille, jotka myös juovat paljon; rikkoutuneiden näyttöjen vakuutus olisi suositeltava suurikokoisten puhelinten käyttäjille.

Vakuutusyhtiöt voivat tunnistaa kohderyhmät tarkalla tietojen analysoinnilla, suorittaa kohdennetun markkinoinnin ja edistää asiaan kuuluvia vakuutustuotteita asiakkaiden yksilöllisten tarpeiden mukaisesti, jotta vältetään potentiaalisten asiakkaiden loukkaaminen lähettämällä massamainoksia.(Big data: boosting insurance development and innovation 2016.)

3 AINESTON ANALYYSI

Tässä työssä käytetään avointen data-aineistojen ”Tieliikenneonnettomuudet vuodelta 2015”, ”Tieliikenneonnettomuudet vuodelta 2016” ja ”Tieliikenneonnettomuudet vuodelta 2017” tietoja. Liikennevirasto kerää vuosittain tieliikenneonnettomuuksiin liittyvää dataa poliisilta saatujen tietojen perusteella ja täydentää ne tilastokeskuksen avustuksella.

Ainesto on toteutettu Datapackage-standardin ohjeiden mukaan. Aineisto sisältää taulukoita CSV-muodossa sekä tietoja onnettomuudesta ja onnettomuuteen osallisista henkilöistä. Aineisto ei sisällä tietoja Ahvenanmaalla sattuneista onnettomuuksista.

Tieliikenneonnettomuus on omaisuusvahinkoja ja/tai henkilövahinkoja aiheuttanut kulkuneuvon liikkumisesta johtunut liikennetapahtuma, jossa on ollut osallisena ainakin yksi liikkuva ajo- taikka kulkuneuvo ja joka on tapahtunut liikenteeseen yleisesti käytetyllä alueella.(Tieliikenneonnettomuudet 2019.)

3.1 Tietojen valmistelu (data preprocessing)

Raakatiedot ovat erittäin herkkiä melulle, puuttuville arvoille ja epäjohdonmukaisuudelle. Niinpä on tärkeää, että nämä tiedot käsitellään ennen kuin niitä louhitaan. Esikäsittely-tiedot ovat olennainen askel tietotehokkuuden parantamiseksi.

Tietojen esikäsittelymenetelmät jaetaan seuraaviin luokkiin:

- Tietojen puhdistus (Data Cleaning)
- Tietojen integrointi (Data Integration)
- Tietojen muuntaminen (Data Transformation)
- Tietojen vähentäminen (Data Reduction) (Alasadi 2017.)

Ensimmäinen tehtävä, jonka tutkija joutuu kohtaamaan, on perustaa työhakemisto ja tuoda tietoja. Kun haluaa tarkastella tuontitietoja, käytetään toimintoa View(), head().

Toinen tehtävä on arvojen ja tyyppitarkastusten selvittäminen. Tarkastellaan tietorakennetta käyttämällä str() -toimintoa. Tietotyypit ovat integer ja Factor.

Kolmas tehtävä on työskennellä merkkijonoilla. Koska ladatussa päivämääräsarjassa oli lukemattomia merkkejä, asetetaan sarakkeiden nimet "Ikä", "Ajokortikä". Saraken nimi ”Nopsuunvas ” vaihdetaan Nopeusrajoitueksi.

Neljäs tehtävä on työskennellä puuttuvien arvojen, tunnistus- ja käsittelymenetelmien avulla. Päivä joukoissa on puuttuvia arvoja, jotka poistetaan na.omit() -toiminnolla.

Viides tehtävä on muuttaa muuttujia: muuntaa, luoda, poistaa. Havaintojen (rivien) valinta tai poistaminen on useimmissa tapauksissa avain onnistuneeseen tietojen valmisteluun ja analysointiin. On tarpeen valita analyysissä käytettävät arvot. Yhdistetään kolme taulukkoa ”tieliikenneonnettomuudet_2015_hlo.csv”, ”tieliikenneonnettomuudet_2016_hlo.csv” ja ”tieliikenneonnettomuudet_2017_hlo.csv” yhteen ja valitaan havaintoja työn helpottamiseksi. Kaikki analyysin valmisteluvaiheet ovat kuvissa 2, 3 ja 4.

```
Sys.getlocale("LC_ALL")
setwd("~/Desktop")
library("stringi", lib.loc="/Library/Frameworks/R.framework/Versions/3.3/Resources/library")
library("stringr", lib.loc="/Library/Frameworks/R.framework/Versions/3.3/Resources/library")
library("tidyr", lib.loc="/Library/Frameworks/R.framework/Versions/3.3/Resources/library")
library("zoo", lib.loc="/Library/Frameworks/R.framework/Versions/3.3/Resources/library")

# Readind data (tieliikenneonnettomuudet_20_hlo), work with NA, subset
tieliikenneonnettomuudet_2017_hlo <- read.csv("tieliikenneonnettomuudet_2017_hlo.csv", sep = ";", check.names = F)
View(tieliikenneonnettomuudet_2017_hlo)
srt(tieliikenneonnettomuudet_2017_hlo)
colnames(tieliikenneonnettomuudet_2017_hlo)[6] <- "Ikä"
colnames(tieliikenneonnettomuudet_2017_hlo)[13] <- "Ajokortikä"
tieliikenneonnettomuudet_2017_hlo_subset <- tieliikenneonnettomuudet_2017_hlo[, c(1, 6, 13)]
tieliikenneonnettomuudet_2017_hlo_subset_without_NA <- na.omit(tieliikenneonnettomuudet_2017_hlo_subset)
View(tieliikenneonnettomuudet_2017_hlo_subset_without_NA)


tieliikenneonnettomuudet_2016_hlo <- read.csv("tieliikenneonnettomuudet_2016_hlo.csv", sep = ";", check.names = F)
View(tieliikenneonnettomuudet_2016_hlo)
srt(tieliikenneonnettomuudet_2016_hlo)
colnames(tieliikenneonnettomuudet_2016_hlo)[6] <- "Ikä"
colnames(tieliikenneonnettomuudet_2016_hlo)[13] <- "Ajokortikä"
tieliikenneonnettomuudet_2016_hlo_subset <- tieliikenneonnettomuudet_2016_hlo[, c(1, 6, 13)]
tieliikenneonnettomuudet_2016_hlo_subset_without_NA <- na.omit(tieliikenneonnettomuudet_2016_hlo_subset)
View(tieliikenneonnettomuudet_2016_hlo_subset_without_NA)

tieliikenneonnettomuudet_2015_hlo <- read.csv("tieliikenneonnettomuudet_2015_hlo.csv", sep = ";", check.names = F)
View(tieliikenneonnettomuudet_2015_hlo)
srt(tieliikenneonnettomuudet_2015_hlo)
colnames(tieliikenneonnettomuudet_2015_hlo)[6] <- "Ikä"
colnames(tieliikenneonnettomuudet_2015_hlo)[13] <- "Ajokortikä"
tieliikenneonnettomuudet_2015_hlo_subset <- tieliikenneonnettomuudet_2015_hlo[, c(1, 6, 13)]
tieliikenneonnettomuudet_2015_hlo_subset_without_NA <- na.omit(tieliikenneonnettomuudet_2015_hlo_subset)
View(tieliikenneonnettomuudet_2015_hlo_subset_without_NA)


#join 3 tables in 1 (tieliikenneonnettomuudet_20XX_hlo)
new_table_tieliikenneonnettomuudet_2017_2016_2015_hlo <- rbind(tieliikenneonnettomuudet_2017_hlo_subset_without_NA,
tieliikenneonnettomuudet_2016_hlo_subset_without_NA, tieliikenneonnettomuudet_2015_hlo_subset_without_NA)
View(new_table_tieliikenneonnettomuudet_2017_2016_2015_hlo)

```

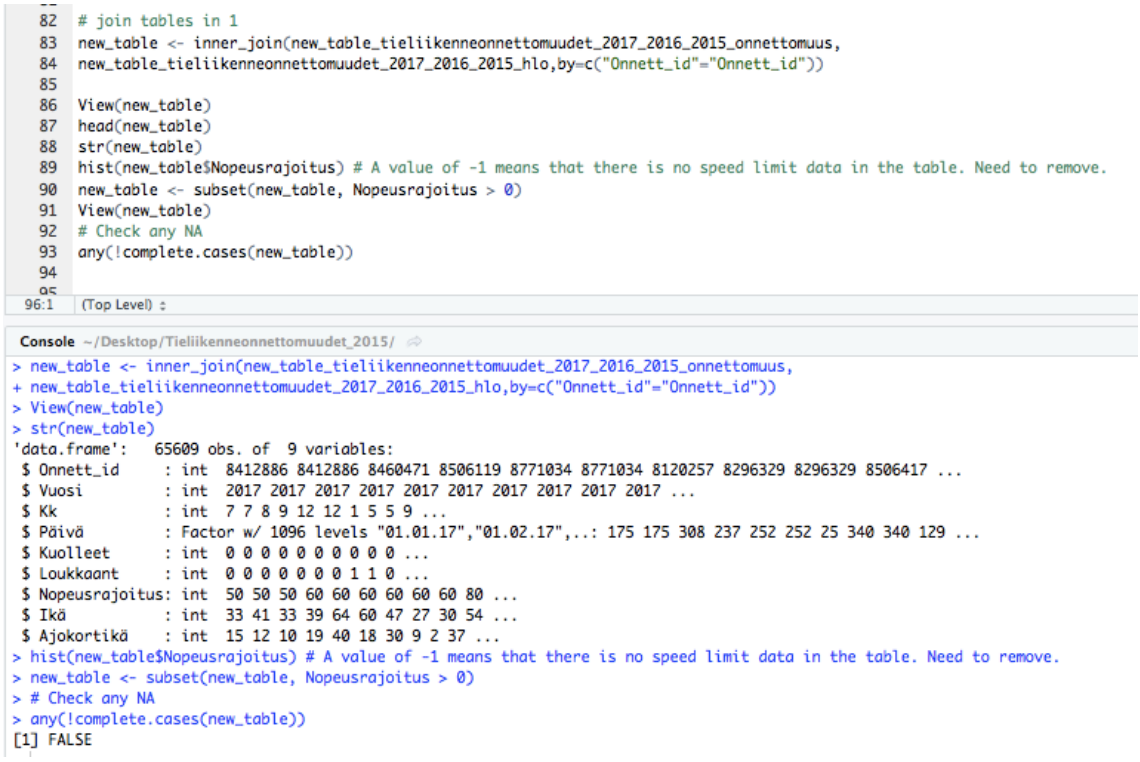
Kuva 2. Tietojen valmistelu.

Yhdistetään kolme taulukkoa ”tieliikenneonnettomuudet_2015_onnettomuus.csv”, ”tieliikenneonnettomuudet_2016_ onnettomuus.csv” ja ”tieliikenneonnettomuudet_2017_ onnettomuus.csv” yhteen.

```
# Readind data (tieliikenneonnettomuudet_20XX_onnettomuus)
tieliikenneonnettomuudet_2017_onnettomuus <- read.csv("tieliikenneonnettomuudet_2017_onnettomuus.csv", sep = ";", check.names = F )
View(tieliikenneonnettomuudet_2017_onnettomuus)
srt(tieliikenneonnettomuudet_2017_onnettomuus)
colnames(tieliikenneonnettomuudet_2017_onnettomuus)[10] <- "Päivä"
which(colnames(tieliikenneonnettomuudet_2017_onnettomuus) == "Nopsuunvas")
colnames(tieliikenneonnettomuudet_2017_onnettomuus)[46] <- "Nopeusrajoitus"
tieliikenneonnettomuudet_2017_onnettomuus_subset <- tieliikenneonnettomuudet_2017_onnettomuus[, c(1, 8:12, 46)]
tieliikenneonnettomuudet_2017_onnettomuus_subset_without_NA <- na.omit(tieliikenneonnettomuudet_2017_onnettomuus_subset)
View(tieliikenneonnettomuudet_2017_onnettomuus_subset_without_NA)


tieliikenneonnettomuudet_2016_onnettomuus <- read.csv("tieliikenneonnettomuudet_2016_onnettomuus.csv", sep = ";", check.names = F )
View(tieliikenneonnettomuudet_2016_onnettomuus)
srt(tieliikenneonnettomuudet_2016_onnettomuus)
colnames(tieliikenneonnettomuudet_2016_onnettomuus)[10] <- "Päivä"
colnames(tieliikenneonnettomuudet_2016_onnettomuus)[46] <- "Nopeusrajoitus"
tieliikenneonnettomuudet_2016_onnettomuus_subset <- tieliikenneonnettomuudet_2016_onnettomuus[, c(1, 8:12, 46)]
tieliikenneonnettomuudet_2016_onnettomuus_subset_without_NA <- na.omit(tieliikenneonnettomuudet_2016_onnettomuus_subset)
View(tieliikenneonnettomuudet_2016_onnettomuus_subset_without_NA)


tieliikenneonnettomuudet_2015_onnettomuus <- read.csv("tieliikenneonnettomuudet_2015_onnettomuus.csv", sep = ";", check.names = F )
View(tieliikenneonnettomuudet_2015_onnettomuus)
srt(tieliikenneonnettomuudet_2015_onnettomuus)
colnames(tieliikenneonnettomuudet_2015_onnettomuus)[10] <- "Päivä"
colnames(tieliikenneonnettomuudet_2015_onnettomuus)[46] <- "Nopeusrajoitus"
tieliikenneonnettomuudet_2015_onnettomuus_subset <- tieliikenneonnettomuudet_2015_onnettomuus[, c(1, 8:12, 46)]
tieliikenneonnettomuudet_2015_onnettomuus_subset_without_NA <- na.omit(tieliikenneonnettomuudet_2015_onnettomuus_subset)
View(tieliikenneonnettomuudet_2015_onnettomuus_subset_without_NA)


# join 3 tables in 1 (tieliikenneonnettomuudet_20XX_onnettomuus)
new_table_tieliikenneonnettomuudet_2017_2016_2015_onnettomuus <- rbind(tieliikenneonnettomuudet_2017_onnettomuus_subset_without_NA,
tieliikenneonnettomuudet_2016_onnettomuus_subset_without_NA, tieliikenneonnettomuudet_2015_onnettomuus_subset_without_NA)
View(new_table_tieliikenneonnettomuudet_2017_2016_2015_onnettomuus)

```

Kuva 3. Tietojen valmistelu, osa 2.

Kaikki arvot "- 1" sarakkeesta "Nopeusraja" poistetaan taulukosta. "-1" tarkoittaa, että nopeusrajoitustietoja ei ole. Käyttämällä "any(complete.cases())" -komentoa tarkistetaan, että taulukossa ei ole puuttuvia arvoja. Tulokset yhdistetään “new_table”-nimellä yhdeksi taulukoksi, jossa on 9 muuttujat.



```
82  # join tables in 1
83  new_table <- inner_join(new_table_tieliikenneonnettomuudet_2017_2016_2015_onnettomuus,
84  new_table_tieliikenneonnettomuudet_2017_2016_2015_hlo,by=c("Onnett_id"="Onnett_id"))
85
86  View(new_table)
87  head(new_table)
88  str(new_table)
89  hist(new_table$Nopeusrajoitus) # A value of -1 means that there is no speed limit data in the table. Need to remove.
90  new_table <- subset(new_table, Nopeusrajoitus > 0)
91  View(new_table)
92  # Check any NA
93  any(!complete.cases(new_table))
94
96:1  (Top Level)
```

```
Console ~/Desktop/Tieliikenneonnettomuudet_2015/
> new_table <- inner_join(new_table_tieliikenneonnettomuudet_2017_2016_2015_onnettomuus,
+ new_table_tieliikenneonnettomuudet_2017_2016_2015_hlo,by=c("Onnett_id"="Onnett_id"))
> View(new_table)
> str(new_table)
'data.frame':   65609 obs. of  9 variables:
 $ Onnett_id     : int  8412886 8412886 8460471 8506119 8771034 8771034 8120257 8296329 8296329 8506417 ...
 $ Vuosi         : int  2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 ...
 $ Kk            : int  7 7 8 9 12 12 1 5 5 9 ...
 $ Päivä         : Factor w/ 1096 levels "01.01.17","01.02.17",..: 175 175 308 237 252 252 25 340 340 129 ...
 $ Kuolleet      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Loukkaant     : int  0 0 0 0 0 0 0 1 1 0 ...
 $ Nopeusrajoitus: int  50 50 50 60 60 60 60 60 60 80 ...
 $ Ikä           : int  33 41 33 39 64 60 47 27 30 54 ...
 $ Ajokortikä    : int  15 12 10 19 40 18 30 9 2 37 ...
> hist(new_table$Nopeusrajoitus) # A value of -1 means that there is no speed limit data in the table. Need to remove.
> new_table <- subset(new_table, Nopeusrajoitus > 0)
> # Check any NA
> any(!complete.cases(new_table))
[1] FALSE
```

Kuva 4. Valmiita tietoja analyysia varten.

Raakatietojen käsittelyn tuloksena tiedot on sijoitettu yhdeksi taulukoksi. Puuttuvia arvoja ei ole ja tiedot ovat valmiita analysointia varten.

3.2 Liikenneonnettomuuksien uhrien lukumäärään vaikuttavat tekijät

Lineaarisessa regressiossa mallitetaan tilannetta, jossa *selitettävä muuttuja* Y riippuu lineaarisesti *selittävistä muuttujista* X_1, X_2 ,... X_p. Sekä selitettävä muuttuja että selitettävät muuttujat ovat *intervalliasteikollisia*. Malli voidaan kirjoittaa muodossa:

$$Y = b_0 + b_1*X_1 + ... + b_p*X_p + E$$

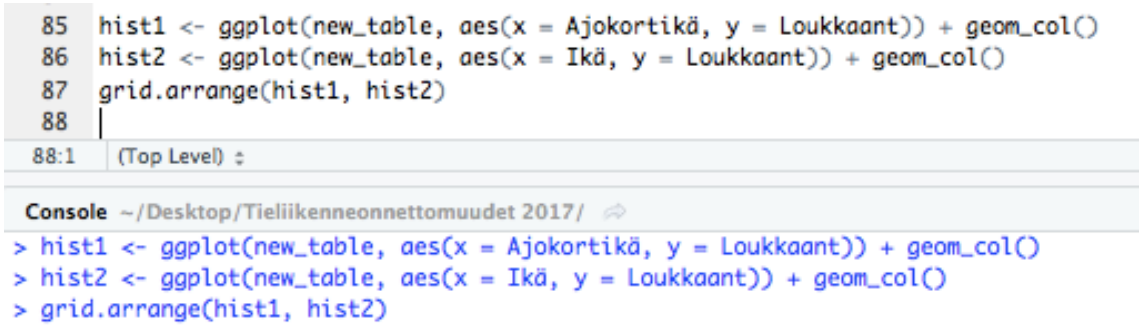
missä b_0, b_1, ..., b_p ovat estimoitavat *regresssiokertoimet* ja muuttuja E on mallin *residuaali*. Muuttujaa

$$PRED\_Y = b_0 + b_1*X_1 + ... + b_p*X_p$$

sanotaan mallin *ennusteeksi*. (A. Nevanlinna, 2002.)

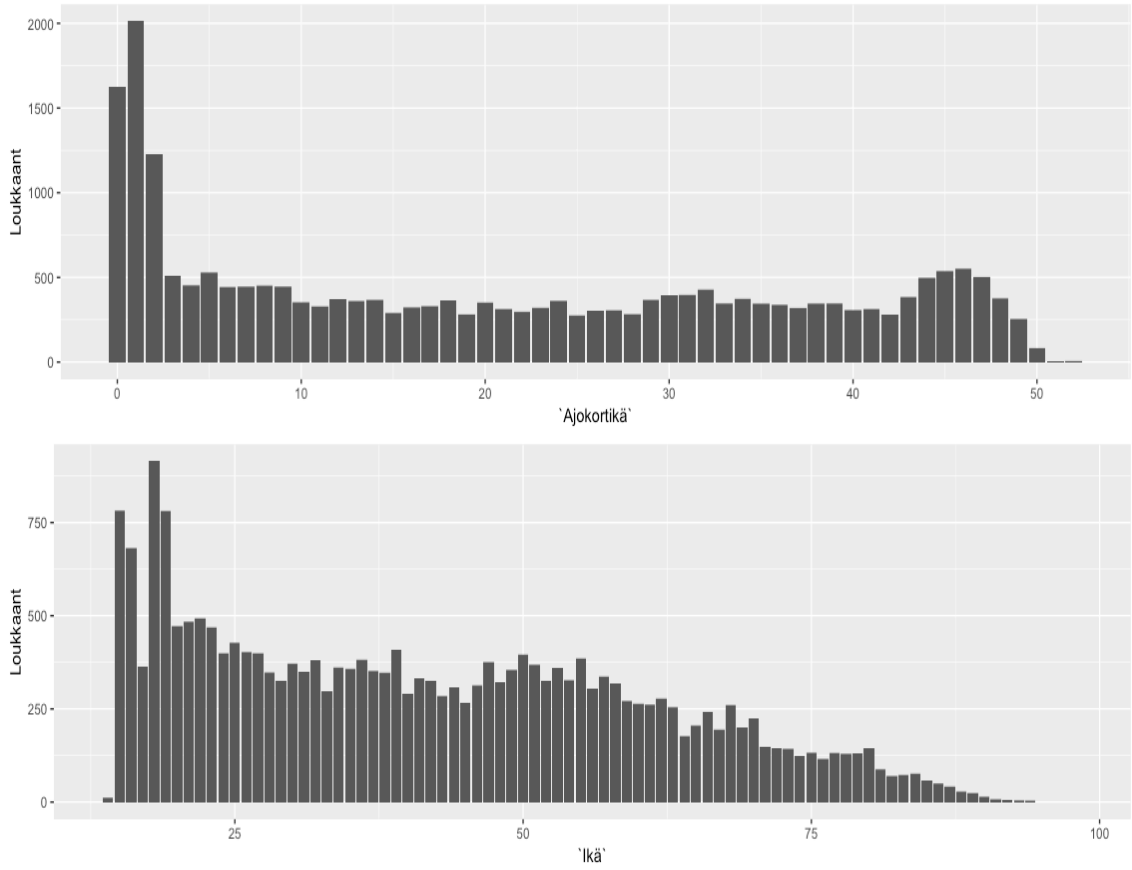
Pyritään selvittämään usean selittävän muuttujan regressioanalyysin avulla, miten kuljettajan ikä ja ajokokemus selittävät liikenneonnettomuuksissa loukkaantuneiden määrä.

Kuvassa 5 on rakennettu ggplot().



Kuva 5. Ggplotin luominen.

Kuvasta 6 voidaan nähdä ilmi, että liikenneonnettomuuksien uhrien lukumäärä on suurempi kuljettajille, joilla on kokemusta enintään 3 vuotta. Sen lisäksi nuorilla on selvästi muita kuljettajia suurempi onnettomuusriski.



Kuva 6. Kuljettajan iän, ajokokemuksen ja uhrien määrän riippuvuus.

Tehdään regressioanalyysi, joka tarkistaa kuljettajan iän ja ajokokemuksen vaikutuksen liikenneonnettomuuksissa loukkaantuneiden ihmisten määrään.

```
> Multiple_Linear_Regression <- lm(Loukkaant ~ Ikä + Ajokortikä, new_table)
> summary(Multiple_Linear_Regression)

Call:
lm(formula = Loukkaant ~ Ikä + Ajokortikä, data = new_table)

Residuals:
   Min     1Q Median     3Q    Max
-0.708 -0.413 -0.333  0.478 11.656

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.421322   0.016444   25.62  < 2e-16 ***
Ikä          0.003942   0.000729    5.41  6.3e-08 ***
Ajokortikä  -0.008946   0.000846  -10.58  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.793 on 29492 degrees of freedom
Multiple R-squared:  0.0084,    Adjusted R-squared:  0.00833
F-statistic:  125 on 2 and 29492 DF,  p-value: <2e-16
```

Kuva 7. Usean selittävän muuttujan regressioanalyysi.

P-arvo - on nollahypoteesin pätevyyden todennäköisyys, jossa todetaan, että riippumattomat muuttujat eivät selitä riippuvan muuttujan dynamiikkaa. Jos p-arvo on kynnysarvon alapuolella (0.05), nollahypoteesi on väärä. Mitä p-arvoa pienempi, sitä parempi.

Tulos: On todettu, että ikä ja ajokokemus vaikuttavat merkittävästi onnettomuuksien määrään, koska p – value 6.3e-08 ja <2e-16, mikä on pienempi kuin 0.05 arvoa. Tällainen pieni arvo merkitsee negatiivista suhdetta:

- mitä vähemmän ajokokemusta on, sitä vahvempi on yhteys tieliikenteessä loukkaantuneiden määrä;
- mitä nuorempi kuljettaja on, sitä vahvempi on yhteys tieliikenteessä loukkaantuneiden määrään.

Tämä viittaa siihen, että kuljettajien onnettomuusriski pienenee ajokokemuksen ja iän karttuessa.

Tarkastellaan R2-lukua (R-Squared), joka on korrelaation rinnakkaiskäsite. Se kuvaa regressioanalyysin selitysvoimaa, eli muuttujien välisen tilastollisen riippuvuussuhteen voimakkuutta. Mitä lähempänä R2 on 1:tä, sitä paremmin regressio kuvaa selittävien ja riippuvien muuttujien välistä suhdetta.

On tärkeää muistaa, että korrelaatio tai R2 eivät merkitse kausaliteettia eli seuraussuhdetta. Näin ollen vaikka muuttujien X ja Y välillä on voimakas tilastollinen riippuvuus regressioanalyysissä ei voi päätellä, että X aiheuttaa aina Y:n tai toisin päin.

Tarkasteltavana olevassa tapauksessa korjattu R^2-luku on pieni (*adjusted* R^2 0.00833). Se antaa syyn päätellä, että onnettomuudessa loukkaantuneiden henkilöiden lukumäärään vaikuttavat myös muut tekijät. Voidaan olettaa, että tämä tekijä voi olla nopeusrajoitus.

Rakennetaan toinen regressiomalli, jossa on kolme ennustajaa: ikä, ajokokemus ja nopeusrajoitus.

```
> Multiple_regression_with_3_predictors <- lm(Loukkaant ~ Ikä + Ajokortikä + Nopeusrajoitus, new_table)
> summary(Multiple_regression_with_3_predictors)

Call:
lm(formula = Loukkaant ~ Ikä + Ajokortikä + Nopeusrajoitus,
    data = new_table)

Residuals:
   Min     1Q Median     3Q    Max
-0.729 -0.413 -0.332  0.476 11.673

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.487929   0.024625   19.81  < 2e-16 ***
Ikä             0.003769   0.000730    5.16  2.4e-07 ***
Ajokortikä     -0.008770   0.000847  -10.36  < 2e-16 ***
Nopeusrajoitus -0.000797   0.000219   -3.63  0.00028 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.793 on 29491 degrees of freedom
Multiple R-squared:  0.00884,   Adjusted R-squared:  0.00874
F-statistic: 87.7 on 3 and 29491 DF,  p-value: <2e-16
```

Kuva 8. Kolmen muuttujan regressioanalyysi.

Tulos: On todettu, kuten edellisessä mallissa, että kuljettajan ikä, ajokokemus ja nopeusrajoitus vaikuttavat onnettomuuksissa loukkaantuneiden lukumäärään, koska p-arvot 2.4e-07, <2e-16 ja 0.0028 vähemmän 0.05 arvoa. Nollahypoteesi on väärä.

Tällainen pieni p-arvo merkitsee negatiivista suhdetta seuraavissa tapauksissa:

- mitä vähemmän ajokokemusta on, sitä vahvempi yhteys tieliikenteessä loukkaantuneiden määrään;
- mitä nuorempi kuljettaja on, sitä vahvempi yhteys tieliikenteessä loukkaantuneiden määrään;
- mitä suurempi nopeusrajoitus, sitä vahvempi yhteys tieliikenteessä loukkaantuneiden määrään.

Korjattu R^2-luku arvo on pieni (*adjusted* R^2 0.00874). Se antaa syyn päätellä, että onnettomuudessa loukkaantuneiden henkilöiden lukumäärään vaikuttavat myös muut tekijät. Voidaan olettaa, että alkoholin ja huumeiden myrkytys voi olla tällainen tekijä.

Verrataan kahta mallia ANOVA -testin avulla. Varianssianalyysia (*analysis of variance* tai *ANOVA*) käytetään tutkittaessa eroavatko kahden tai useamman ryhmän keskiarvot tilastollisesti merkitsevästi toisistaan.

```
> # Comparing models
> anova(Multiple_regression_with_3_predictors, Multiple_Linear_Regression)
Analysis of Variance Table

Model 1: Loukkaant ~ Ikä + Ajokortikä + Nopeusrajoitus
Model 2: Loukkaant ~ Ikä + Ajokortikä
  Res.Df   RSS Df Sum of Sq    F  Pr(>F)
1  29491 18548
2  29492 18556 -1      -8.3 13.2 0.00028 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Kuva 9. Regressiomallien vertailu.

Tulos: ANOVA -testin avulla on todettu, että paras malli on regressiomalli, jossa on 2 ennustajaa (kuljettajan ikä ja ajokokemus).

Tehdään ennusteen parhaasta mallista. Esimerkiksi, kuinka vaikuttavat uhrien määrään kuljettajan 19 vuoden ikä ja yhden vuoden ajokokemus, kuljettajan 30 vuoden ikä ja 5 vuoden ajokokemus, kuljettajan 50 vuoden ikä ja 19 vuoden ajokokemus.

```
> Data_for_prediction <- data.frame(
+   Ikä = c(19, 30, 50),
+   Ajokortikä = c(1, 9, 19)
+ )
> Prediction_Multiple_Linear_Regression <- predict(Multiple_Linear_Regression, newdata = Data_for_prediction,
+                                                  interval = "confidence", level = 0.95)
> Prediction_Multiple_Linear_Regression
    fit   lwr   upr
1 0.487 0.470 0.504
2 0.459 0.446 0.472
3 0.448 0.430 0.466
```

Kuva 10. Confidence intervals.

Tulos: Regressioyhtälön ennustaminen on vastaavan arvon x korvaaminen regressioyhtälössä. Tätä ennustetta kutsutaan pisteeksi. Se ei ole tarkka, joten sitä täydennetään standardivirheen laskennalla, minkä seurauksena saadaan ennustetun arvon väliarvio.

Yllä olevassa kuvassa on ennusteen luottamusväli (istuntojakso). 95 % ennustejaksoista on merkitty parametreilla lwr ja upr. 95 %: n luottamuksella voidaan väittää, että 19-vuotias kuljettaja, jolla on 1 vuoden ajokokemus 0–1 uhria onnettomuudessa. Samankaltaisia tuloksia saatiin 30-vuotiaasta kuljettajasta, jolla oli 9 vuoden kokemus, ja 50-vuotiasta kuljettajasta, jolla oli 19 vuoden ajokokemus, mutta pienemmät luottamusvälit. Tämä tarkoittaa sitä, että onnettomuuden loukkaantumisen todennäköisyys on pienempi, jos kuljettaja on aikuinen ja kokenut kuljettaja.

3.3 Aikasarjat liikenneonnettomuuksien määrän ennustamiseksi vuonna 2018

Tietoja liikenneonnettomuuksista vuonna 2018 ei ole vielä julkaistu Internetissä. Aikasarjojen avulla on mahdollista ennustaa onnettomuuksien määrää esimerkiksi vuoden ajan.

Aikasarjojen ennustamiseksi käytetään parametreja kuukausittaisten onnettomuuksien ja uhrien lukumäärä.

	Kk_Vuosi	Loukkaant
1	2015-01-01	455
2	2015-02-01	327
3	2015-03-01	396
4	2015-04-01	377
5	2015-05-01	585
6	2015-06-01	586
7	2015-07-01	683
8	2015-08-01	737
9	2015-09-01	574
10	2015-10-01	511
11	2015-11-01	551
12	2015-12-01	546
13	2016-01-01	436
14	2016-02-01	340
15	2016-03-01	265
16	2016-04-01	399

Showing 1 to 17 of 36 entries

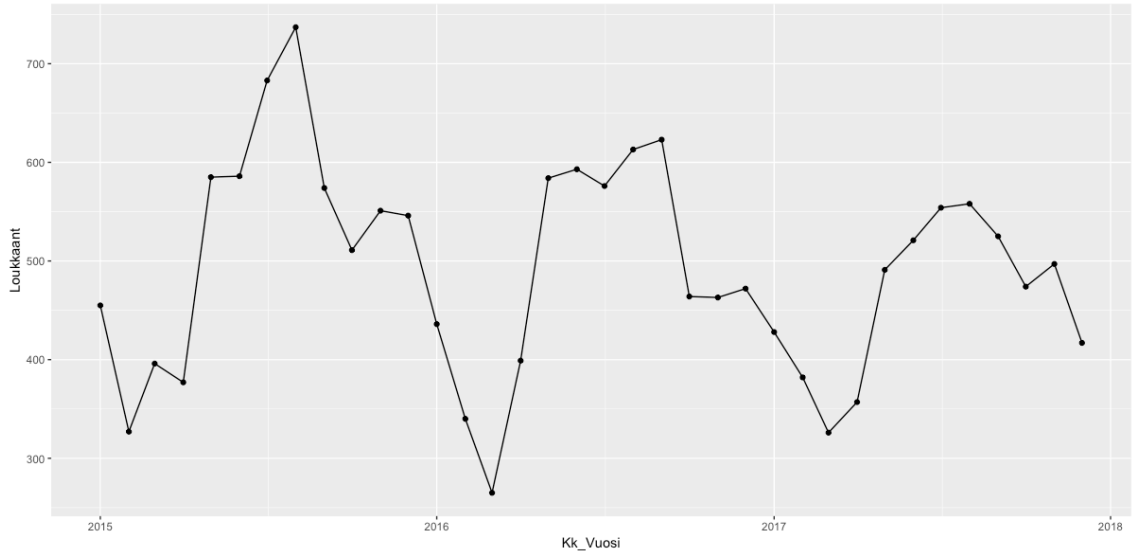
Console ~/Desktop/Tieliikenneonnettomuus_2016/

```
> View(data_for_TS)
> str(data_for_TS)
'data.frame':   36 obs. of  2 variables:
 $ Kk_Vuosi : Date, format: "2015-01-01" ...
 $ Loukkaant: int  455 327 396 377 585 586 683 737 574 511 ...
```

Kuva 11. Data set for Time Series Analysis.

Kuvan 11 datatyypissä nähdään parametrit Kk_Vuosi on päivämäärä -datatyypi ja Loukkaant on integer.

Ggplot(data_for_TS, aes (Kk_Vuosi, Loukkaant)) + geom_line() + geom_point() avulla tarkastelemme onnettomuuksien määrää 1.2015–12.2017.



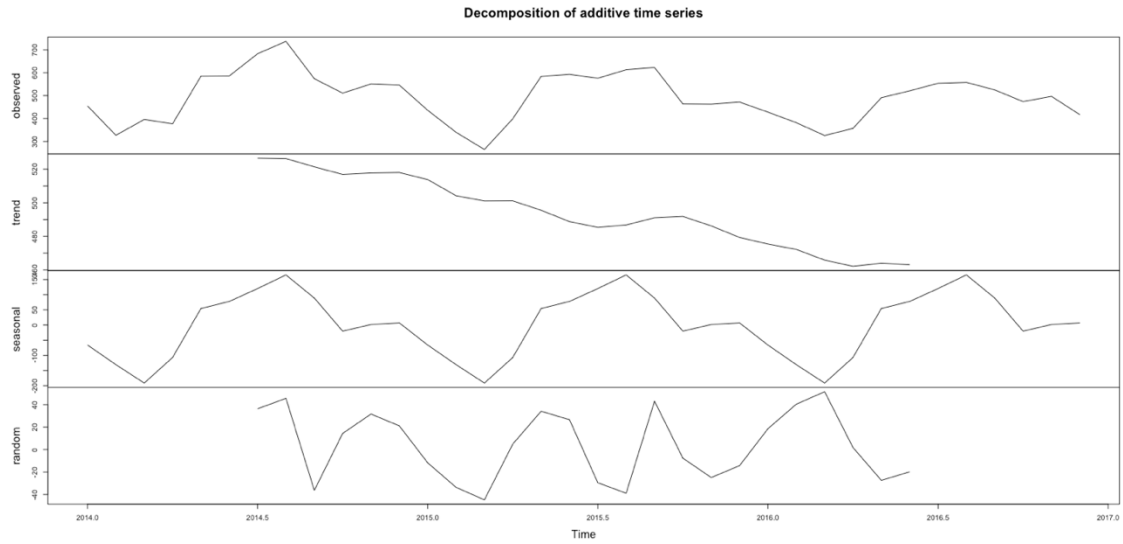
Kuva 12. Tieliikenteessä loukkaantuneet vuonna 2015 – 2017 Suomessa.

Rakennetaan aikasarja. Funktio ts() ottaa numeerisen vektorin, aloitusajan ja mittaustaajuuden. Tässä nämä arvot ovat uhrien lukumäärä, 2015 (vuosi, jona mittaukset alkavat), 1 (kuukausi, jona mittaukset alkavat) ja 12:n taajuus (kuukaudet vuodessa).

```
> dec
$x
     Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
2015 455 327 396 377 585 586 683 737 574 511 551 546
2016 436 340 265 399 584 593 576 613 623 464 463 472
2017 428 382 326 357 491 521 554 558 525 474 497 417

$seasonal
            Jan         Feb         Mar         Apr         May         Jun         Jul         Aug         Sep
2015  -65.965278 -130.548611 -191.298611 -107.006944   54.388889   77.701389  120.201389  165.118056   88.930556
2016  -65.965278 -130.548611 -191.298611 -107.006944   54.388889   77.701389  120.201389  165.118056   88.930556
2017  -65.965278 -130.548611 -191.298611 -107.006944   54.388889   77.701389  120.201389  165.118056   88.930556
            Oct         Nov         Dec
2015  -20.194444    1.680556    6.993056
2016  -20.194444    1.680556    6.993056
2017  -20.194444    1.680556    6.993056

$trend
          Jan      Feb      Mar      Apr      May      Jun      Jul      Aug      Sep      Oct      Nov      Dec
2015       NA       NA       NA       NA       NA       NA 526.5417 526.2917 521.3750 516.8333 517.7083 517.9583
2016 513.7917 504.1667 501.0417 501.1250 495.5000 488.7500 485.3333 486.7500 491.0417 491.8333 486.2083 479.3333
2017 475.4167 472.2083 465.8333 462.1667 464.0000 463.1250       NA       NA       NA       NA       NA       NA

$random
            Jan         Feb         Mar         Apr         May         Jun         Jul         Aug         Sep         Oct
2015         NA          NA          NA          NA          NA          NA  36.256944  45.590278 -36.305556  14.361111
2016 -11.826389 -33.618056 -44.743056   4.881944  34.111111  26.548611 -29.534722 -38.868056  43.027778  -7.638889
2017  18.548611  40.340278  51.465278   1.840278 -27.388889 -19.826389          NA          NA          NA          NA
            Nov         Dec
2015  31.611111  21.048611
2016 -24.888889 -14.326389
2017         NA          NA

$figure
 [1]  -65.965278 -130.548611 -191.298611 -107.006944   54.388889   77.701389  120.201389  165.118056   88.930556
[10]  -20.194444    1.680556    6.993056

$type
[1] "additive"

attr(,"class")
[1] "decomposed.ts"
```

Kuva 13. Time Series structure.

Kuten liitteenä olevista käsikirjoituksista ($ Seasonal) ilmenee, Suomessa onnettomuudet ovat selvästi kausiluonteisia. Voimme myös visualisoida tietosarjojen hajoamisen. Kaaviona se näyttää seuraavanlaiselta (kuva 14):



Kuva 14. Aikasarjan komponentit.

Kuvan 14 kaavio esittää alkuperäisen aikasarjan (ylhäältä), arvioidun trendikomponentin (toinen ylhäältä), arvioidun kausikomponentin (kolmas ylhäältä) ja arvioidun epäsäännöllisen komponentin (pohja).

Nähdään, että arvioitu trendikomponentti osoittaa pysyvää laskua 1.2015–12.2017. Kaaviosta voimme päätellä kausiluonteisuuden.

Teemme 2 testiä paikallisuuden suhteen.

Kiinteässä prosessissa on ominaisuus, että keskiarvo, varianssi ja autokorrelaatiorakenne eivät muutu ajan mittaan. Pysyvyys voidaan määritellä tarkasti matemaattisilla termeillä, mutta tarkoituksessamme tarkoitamme tasaista etsintäsarjaa, jossa ei ole trendiä, jatkuvaa vaihtelua ajan mittaan, jatkuvaa autokorrelaatiorakennetta ajan myötä eikä jaksollisia vaihteluja (kausivaihtelu).

Ensin tarkistetaan Dickey-Fullerin testi epäyhtenäisen aikasarjan x yksikköjuuren nollahypoteesille (vastaavasti x on ei-stationaarinen aikasarja). Toiseksi tarkistetaan Kwiatkowski-Phillips-Schmidt-Shin (KPSS) -testi nollatapahtuman suhteen, että x on taso tai suuntaus pysyvä.

```
292  adf.test(TS, alternative="stationary")
293  kpss.test(TS, null="Trend")
294
295
294:1  (Top Level) ÷

Console ~/Desktop/Tieliikenneonnettomuudet_2015/

> adf.test(TS, alternative="stationary")

        Augmented Dickey-Fuller Test

data:  TS
Dickey-Fuller = -3.9473, Lag order = 3, p-value = 0.02284
alternative hypothesis: stationary

> kpss.test(TS, null="Trend")

        KPSS Test for Trend Stationarity

data:  TS
KPSS Trend = 0.064172, Truncation lag parameter = 1, p-value = 0.1

Warning message:
In kpss.test(TS, null = "Trend") : p-value greater than printed p-value
```

Kuva 15. Stationarity test.

Tulos: P-arvon arvo adf.testissä on 0,02284, joka sallii hylätä sarjan nolla-hypoteesin.

P-arvon arvo kpss.testissä on 0,1, jonka avulla voimme pitää nollahypoteesin olevan totta.

Molemmat testit vahvistivat aikasarjojen pysyvyyden.

Ennustemallin valitseminen

Aiomme kokeilla kolmea eri ennustusmenetelmää. Yleensä se on paras suorituskyky.

Malli 1: Exponential State Smoothing

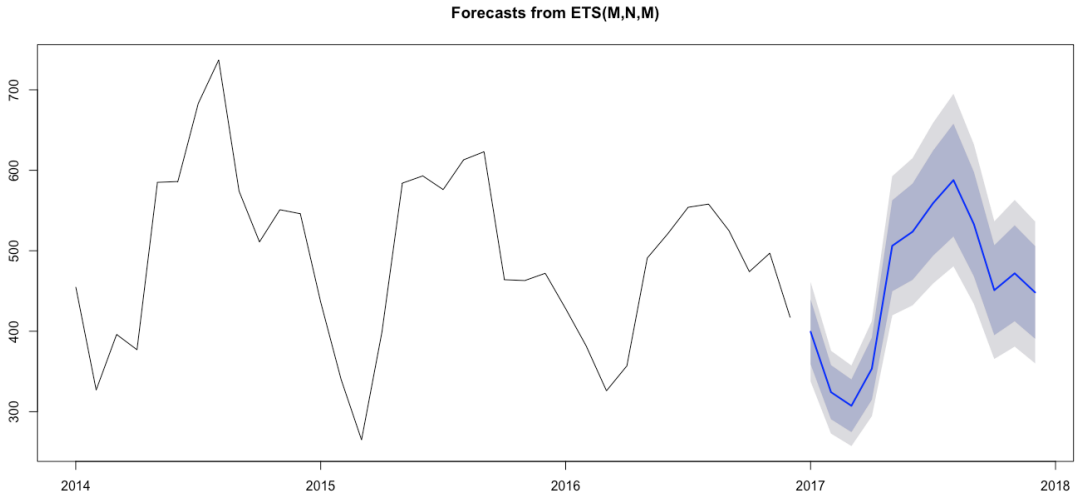
Paketin ets () -funktio sopii eksponentiaalisen tasoitusmallin (ETS) malleihin. Tämä toiminto optimoi automaattisesti mallin parametrit.

Tehdään ennuste seuraaville 12 kuukaudelle.

```
> #1. Forecast
> library(forecast)
> m_ets <- ets(TS)
> f_ets <- forecast(m_ets, h = 12)
> f_ets
         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
Jan 2018       399.3752 359.1557 439.5948 337.8647 460.8857
Feb 2018       324.2841 290.7149 357.8533 272.9445 375.6238
Mar 2018       307.3674 274.7077 340.0271 257.4188 357.3161
Apr 2018       353.4269 314.9294 391.9245 294.5500 412.3038
May 2018       506.1555 449.7020 562.6089 419.8174 592.4936
Jun 2018       523.7210 463.9736 583.4684 432.3453 615.0967
Jul 2018       559.0954 493.9182 624.2727 459.4155 658.7754
Aug 2018       587.7867 517.8292 657.7443 480.7959 694.7775
Sep 2018       532.9987 468.2859 597.7115 434.0290 631.9684
Oct 2018       450.9917 395.1764 506.8070 365.6296 536.3538
Nov 2018       471.9741 412.4738 531.4744 380.9762 562.9720
Dec 2018       448.2025 390.6839 505.7212 360.2353 536.1698
> plot(f_ets)
>
```

Kuva 16. Model Exponential State Smoothing.

Yllä olevassa kuvassa näet ennusteen luottamusvälin (confidence interval). 95% ennustejaksoista on merkitty parametreilla lwr ja upr.



Kuva 17. Prediction of the number of accidents in 2018.

Ennuste esitetään sinisenä, ja harmaa alue edustaa 95%: n luottamusväliä. Katsomalla näemme, että ennuste vastaa suurelta osin tietojen historiallista mallia.

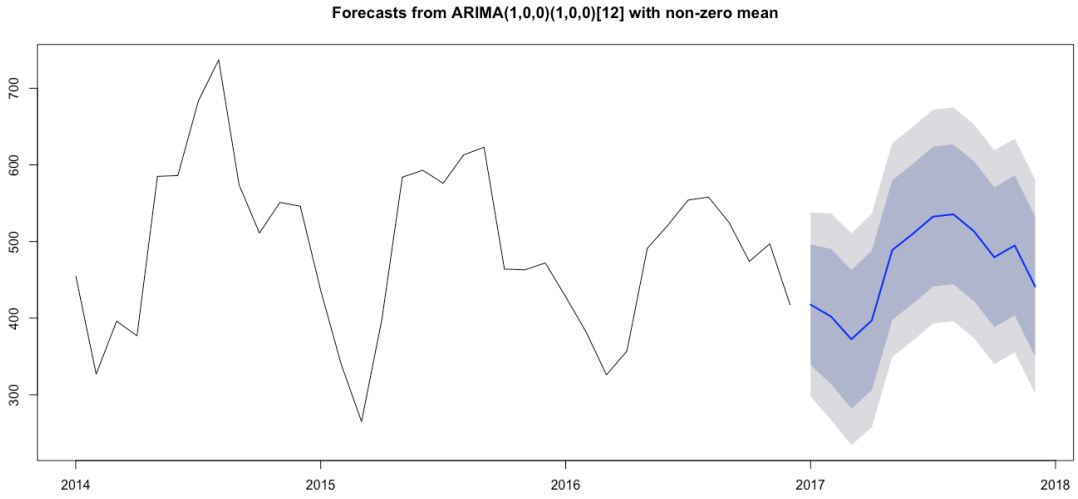
Malli 2: ARIMA

Auto.arima () -toiminto tarjoaa toisen mallintamismenetelmän. Auto.arima () -toiminto etsii automaattisesti parhaan mallin ja optimoi parametrit.

```
> #2. Auto ARima
> m_aa <-  auto.arima(TS)
> f_aa <- forecast(m_aa, h=12)
> f_aa
         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
Jan 2018       417.6381 339.1659 496.1102 297.6253 537.6508
Feb 2018       402.1366 314.1165 490.1566 267.5215 536.7516
Mar 2018       372.4217 282.1006 462.7428 234.2875 510.5559
Apr 2018       397.1172 306.2116 488.0228 258.0890 536.1454
May 2018       488.8118 397.7558 579.8677 349.5538 628.0698
Jun 2018       509.9103 418.8156 601.0050 370.5930 649.2276
Jul 2018       532.5156 441.4109 623.6203 393.1830 671.8482
Aug 2018       535.4559 444.3486 626.5632 396.1194 674.7925
Sep 2018       513.5018 422.3939 604.6098 374.1643 652.8394
Oct 2018       479.4342 388.3261 570.5423 340.0963 618.7720
Nov 2018       494.8631 403.7550 585.9713 355.5252 634.2010
Dec 2018       441.3347 350.2265 532.4429 301.9968 580.6726
> plot(f_aa)
```

Kuva 18. Malli ARIMA.

Yllä olevassa kuvassa näemme ennusteen luottamusvälin (confidence interval). 95% ennustejaksoista on merkitty parametreilla lwr ja upr.

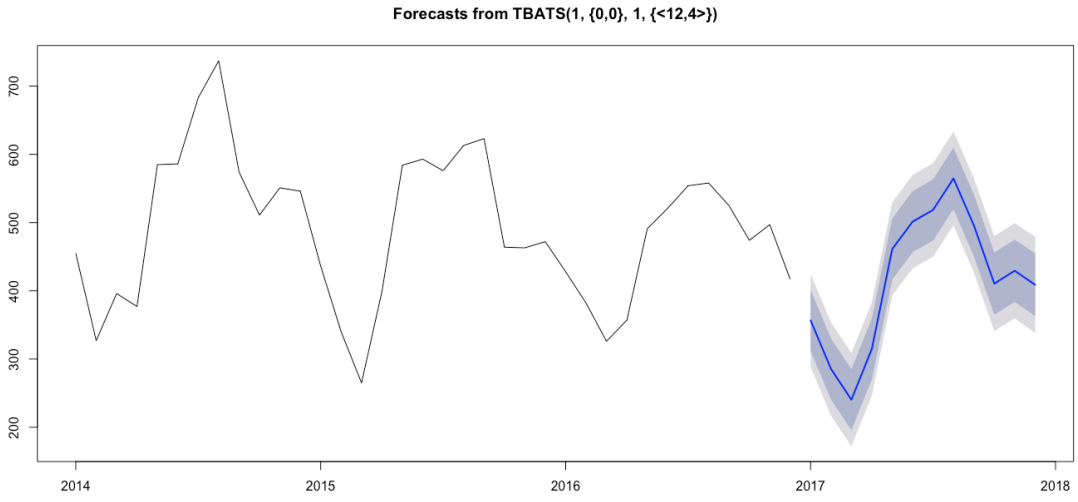


Kuva 19. Ennuste onnettomuuksien määrästä vuonna 2018.

Rakennetaan kolmas malli ja verrataan tuloksia.

```
> #3. Model 3: TBATS
> m_tbats <- tbats(TS)
> f_tbats <- forecast(m_tbats, h=12)
> f_tbats
         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
Jan 2018       356.7128 312.2552 401.1704 288.7208 424.7048
Feb 2018       285.4437 240.9533 329.9342 217.4015 353.4860
Mar 2018       240.3775 195.8803 284.8747 172.3249 308.4302
Apr 2018       314.7867 270.2844 359.2890 246.7264 382.8471
May 2018       461.3068 416.7861 505.8274 393.2183 529.3952
Jun 2018       501.3698 456.7284 546.0111 433.0967 569.6428
Jul 2018       518.5541 473.8072 563.3010 450.1196 586.9885
Aug 2018       564.6702 519.8495 609.4909 496.1228 633.2176
Sep 2018       495.8052 450.7177 540.8927 426.8498 564.7606
Oct 2018       410.6005 365.2361 455.9649 341.2216 479.9794
Nov 2018       429.3993 383.8705 474.9281 359.7691 499.0295
Dec 2018       408.8647 362.8380 454.8914 338.4730 479.2565
> plot(f_tbats)
```

Kuva 20. TBATS forecast.

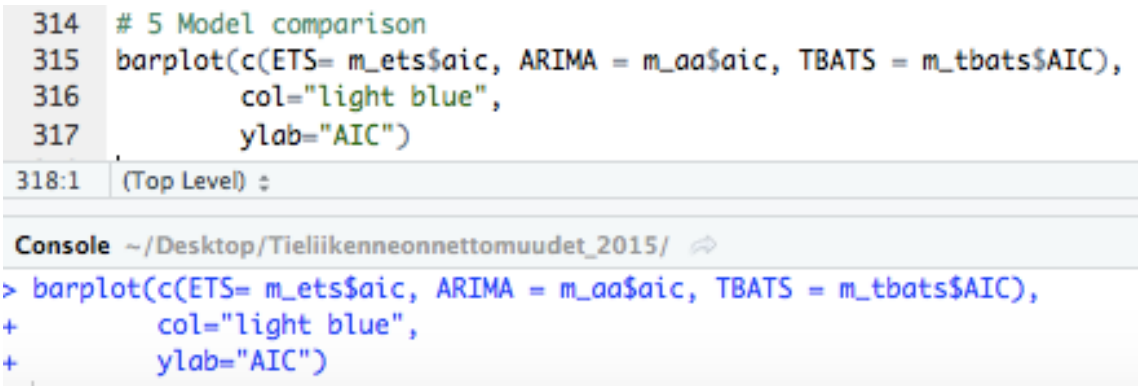


Kuva 21. Ennuste onnettomuuksien määrästä vuonna 2018.

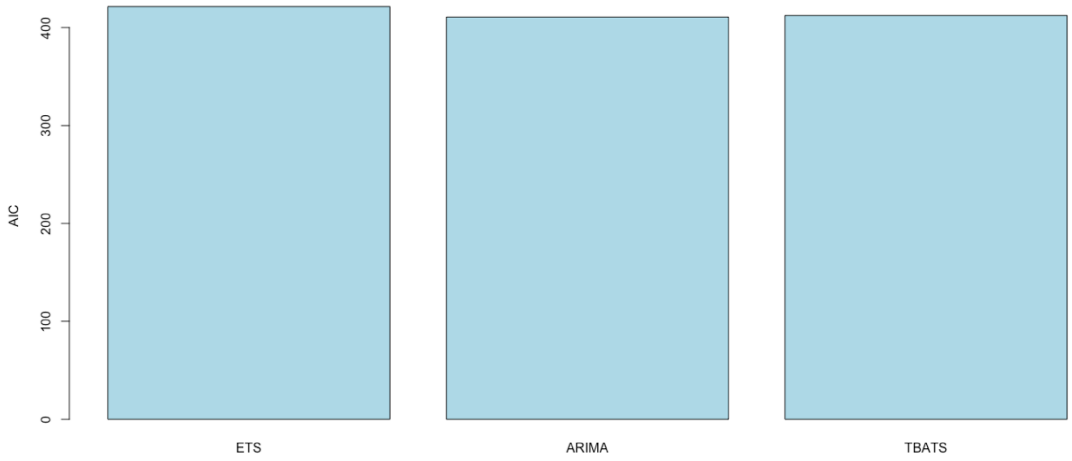
Kuva on ennuste, onnettomuuksien määrästä vuonna 2018. Nyt meillä on kolme mallia, jotka näyttävät antaavan kohtuulliset ennusteet. Verrataan niitä nähdäksemme, mikä toimii parhaiten.

Mallin vertailu

Käytetään AIC: ta vertaamaan eri malleja. AIC on yleinen menetelmä sen määrittämiseksi, kuinka hyvin malli sopii dataan, ja samalla rangaista monimutkaisempia malleja. Pienin AIC-malli on parhaiten sopiva malli.



Kuva 22. Mallien vertailu.



Kuva 23. Mallien visuaalinen vertailu.

Kuten edellä olevasta kaaviosta voidaan nähdää, parhaat mallit ovat auto.arima ja TBAST. Selvitetään, mikä malli on parempi MAPE:n avulla, joka tulkitaan keskimääräisenä absoluuttisena virheenä prosentteina (absoluuttinen virhe, MAPE). MAPE ilmaistaan prosentteina. Mitä alhaisempi tulos, sitä parempi malli.

```
319
320  accuracy(f_ets)
321  accuracy(f_aa)
322  accuracy(f_tbats)
323
323:1  (Top Level) ÷
Console ~/Desktop/Tieliikenneonnettomuudet_2015/
> accuracy(f_ets)
                    ME     RMSE      MAE       MPE     MAPE      MASE        ACF1
Training set -5.464134 35.47228 30.5284 -1.89793 6.678614 0.5751033 0.003933129
> accuracy(f_aa)
                    ME     RMSE      MAE       MPE     MAPE      MASE      ACF1
Training set -4.902254 58.6253 46.50506 -2.990188 10.17181 0.8760764 0.0619822
> accuracy(f_tbats)
                    ME     RMSE      MAE       MPE     MAPE      MASE        ACF1
Training set -2.670556 34.69043 28.88701 -1.133675 6.545434 0.5441822 0.08159899
>
```

Kuva 24. Accuracy test.

Tulos: Kuten voidaan havaita MAPE:n analyysistä, jonka minimiarvo on 6,545434 f_tbats-mallille. Niinpä paras ennustemalli on f_tbats.

4 LOPUKSI

Tässä opinnäytetyössä on käsitelty Big dataa ja sen roolia liiketoiminnassa. Erityinen paikka annettiin Big datan eduille vakuutusalalla. On todettu, että menneisyyden analysointi voi ennustaa tulevaisuutta. Se auttaa yrityksiä minimoimaan kustannukset, kasvattamaan voittoja ja ymmärtämään asiakkaiden käyttäytymistä.

Käytännön tarkoituksiin tehtiin analyysi liikenneonnettomuuksista Suomessa 1.2015–12.2017.

Koko opinnäytetyö on tehty RStudion avulla. Tutkimuksen aikana esitettiin R-taitoja, mukaan lukien raakatietojen käsittely, puuttuvien arvojen poistaminen, tietojen visualisointi ja saatujen tulosten analysointi.

Tutkimuksen aikana rakennettiin kaksi regressiomallia. Ensimmäinen analyysi tutkii kuljettajan iän ja ajokokemuksen vaikutusta uhrien lukumäärään onnettomuudessa. Toisessa mallissa tutkittiin kuljettajan iän, ajokokemuksen ja nopeusrajoitusten vaikutusta uhrien lukumäärään onnettomuudessa. Paras malli on ensimmäinen malli.

Regressioanalyysin mukaan ajokokemuksella on vahva vaikutus uhrien määrään, mutta se ei ole ainoa merkittävä tekijä. Liikenneonnettomuudet vaikuttavat paljon nuorempiin kuljettajiin, joilla on vähän ajokokemusta. Ikääntyneihin kuljettajiin tämä vaikutus vähenee. Nämä tiedot vahvistavat nykyiset vakuutussäännöt, joissa nuorille kuljettajille, joilla on vähän ajokokemusta, autovakuutus on kalliimpaa kuin kokeneen kuljettajan.

Nopeusrajoitusten ja uhrien lukumäärän suhdetta regressiomalleja verrattaessa pidettiin heikkona.

Analyysin mukaan molemmat regressiomallit eivät selitä täys syy-yhteyttä onnettomuuden uhrien lukumäärään onnettomuudessa. Voidaan olettaa, että uhrien määrään vaikuttavat muut tekijät, kuten huumeiden ja alkoholin myrkytys. Näitä tekijöitä ei tutkittu tietojen puutteen vuoksi.

Regressioanalyysin aikana luotiin luottamusväliä ennustetulle uhrien määrälle 19-vuotiaalla kuljettajalla, jolla oli yhden vuoden ajokokemus, 30-vuotias kuljettaja, jolla on yhdeksän vuoden kokemus, 50-vuotias kuljettaja, jolla on 19 vuoden kokemus. Ennusteiden mukaan luottamusväli laskee (uhrien määrä pienenee), kun ajokokemus ja kuljettajan ikä kasvaavat. Niinpä, käyttämällä regressioanalyysiä ja rakentamalla

luottamusvälejä ennustetun uhrien lukumäärän kanssa, voidaan päätellä, että ajokokemus ja kuljettajan ikä vaikuttavat suuresti uhrien määrään.

Sen lisäksi opinnäytetyön aikana käytettiin datan louhintamenetelmiä liikenneonnettomuuksien määrän ennustamiseksi vuonna 2018. Tulosten mukaan onnettomuuksien määrä vuonna 2018 säilyy vuoden 2017 tasolla, mutta vähemmän kuin vuonna 2015 ja 2016.

Valitettavasti opinnäytetyön laatimishetkellä Suomessa ei ole vielä julkaistu tietoa liikenneonnettomuuksista vuonna 2018 eikä saatuja tuloksia voitu verrata. Joka tapauksessa päätehtävä saatiin onnistuneesti päätökseen: tilastotietojen ja tiedon louhinnan menetelmien avulla analysoitiin Big dataa ja ennustettiin tulevia arvoja. Tämä tieto auttaa yrityksiä kasvamaan, ja erityisesti vakuutusyhtiöitä kohdentamaan oikeanhintaiset vakuutukset segmentoidulle asiakkaille.

LÄHTEET

Ada, E. 2018. *How Big data and Machine learning are reshaping the insurance industry*. Viitattu 24.5.2019. https://www.digitaldoughnut.com/articles/2018/september/big-data-and-machine-learning-reshapes-insurance.

Alasadi, S. 2017. *Review of Data Preprocessing Techniques in Data Mining.* Saatavissa https://www.researchgate.net/profile/Suad_Alasadi/publication/320161439_Review_of_Data_Preprocessing_Techniques_in_Data_Mining/links/59d143d64585150177f3d15b/Review-of-Data-Preprocessing-Techniques-in-Data-Mining.pdf.

5 Benefits: competitive advantages of Big Data in business, 2017. Viitattu 24.5.2019. https://www.newgenapps.com/blog/importance-benefits-competitive-advantage-big-data

Big data: boosting insurance development and innovation, 2016. Viitattu 24.5.2019. https://www.swissre.com/china/big_data_boosting_insurance_development_and_innovation.html

KnowledgeHut Editor, 2016. *Types of Big Data*. Viitattu 24.5.2019. https://www.knowledgehut.com/blog/big-data/types-of-big-data

S. Mallon, 2018. *Predictive modelling and Bif data are insurance industry powerhouses.* Viitattu 24.5.2019. https://www.smartdatacollective.com/predictive-modeling-and-big-data-are-insurance-industry-powerhouses/

Marr, B., 2015, s. 15. *Big Data : Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance.*

Minelli, M. & Chambers, M. 2012. *Big Data, Big Analytics: Emerginging Business Intelligence and Analytic Trends for Today`s Businesses.*

Nevanlinna, A., 2002. *Regressioanalyysi.* Viitattu 24.5.2019. http://www.helsinki.fi/~aploper/spssjatko/regressio/regressio.html

Salo, I., 2014, s. 26. *Big data & pilvipalvelut*. Docendo.

Salo, S., 2013, s. 21 – 22. *Big data tiedon vallankumous.* Docendo.

Semi-structured data, 2019. *Wikipedia.* Viitattu 24.5.2019. https://en.wikipedia.org/wiki/Semi-structured_data

Tieliikenneonnettomuudet, 2019. Avoindata.fi. Viitattu 24.5.2019. https://www.avoindata.fi/data/fi/dataset/tieliikenneonnettomuudet