Sissejuhatus andmeteadusesse projekti kirjeldus (tiim W5)

Pealkiri: F1 Analysis (F1 Analüüs)

Tiimiliikmed:
Peeter Paal
Aveli Klaos
Siim Tanel Laisaar

Ülesanne 1

GitHub-i link: https://github.com/PeeterPaal/ids-f1project.
Repositoorium on avalik, seega ligipääsetav ka kõigi juhendajate jaoks.

Ülesanne 2

Äriliste eesmärkide identifitseerimine

Taust

Järgnev projekti kirjeldus on kirjutatud nii nagu projekti läbi viimist oleks palunud *Formula One Group*, mis tegeleb Vormel 1 reklaamimisega ning omab Vormel 1 kommertsõigusi.
Meie klient, *Formula One Group*, tahab kasvatada huvi Vormel 1 vastu. Huvi kasvu soovitakse nii vaatajate kasvu kui ka sponsorite liitumise nurga alt. Uute sponsorite liitumine aitab spordist tulenevat tulu suurendada. Klient soovib parendada ka Vormel 1 meelelahutuslikku väärtust. Selleks soovitakse leida spordi reeglites puudulikke aspekte, mida muuta. Meelelahutusliku väärtuse tõus aitab tagada sponsorite ning vaatajate püsivuse, mis garanteerib spordist tuleneva ühtlase tulu.

Ärilised eesmärgid

Klient soovib projekti käigus loodud andmete visualiseerimise abil tekitada huvi uute võimalike spordi jälgijate ja sponsorite seas. Huvi tõusu all nõutakse, et vaatajaskonna arvukuse tõus ning liituvate sponsorite arv oleks võrreldes varasemate aastatega kõrgem. Samuti soovitakse leida Vormel 1 reeglite seas kitsaskohti, mida tuleks muuta, et sport oleks jälgijatele huvitavam vaadata ning sponsoritele ahvatlevam. Soovitakse leida iga *Grand Prix'ga* kaasneva kvalifikatsiooni ja sõidu juures elemente, mis mängivad sõidu lõpptulemuse juures liiga suurt rolli ning ka neid, mis mängivad liiga väikest rolli. Selle põhjal on võimalik seonduvaid reegleid muuta, mis aitavad spordis osalejate taset ühtlustada, mis parandab spordi meelelahutuslikku väärtust. Meelelahutusliku väärtuse tõusu määratletakse nii uute vaatajate arvu järgi kui ka selle järgi, et lahkuvate vaatajate arv oleks väiksem kui varem.

Ärilise edukuse kriteerium

Kuna viimasel kahel aastal on vaatajaskonna arvukuse tõus olnud ligikaudu 10% aastas, siis edukuse kriteeriumiks loetakse seda, kui vaatajaskonna arvukuse tõus on kuni kahe aasta pärast vähemalt 15% aastas või rohkem. Siiani on vaatajaskonna arvukuse tõus toonud kaasa ka tulu tõusu, seega on see *Formula One Group*-ile äriliselt kasulik. Teiseks edukuse kriteeriumiks loetakse seda, kui suudame ennustamismudeli põhjal leida vähemalt ühe mõjuva kitsaskoha (mõjuvuse üle otsustab FIA ehk *Federation Internationale de l'Automobile*).

Kolmandaks edukuse kriteeriumiks loetakse seda, kui Vormel 1 suursponsorite arv tõuseb järgneva kolme aasta jooksul 2 võrra. Neljandaks edukuse kriteeriumiks on see, et järgmise kolme aasta jooksul lahkuvate vaatajate arv väheneks aastast aastasse.

Situatsiooni analüüsimine

Olemasolevad ressursid

Projekti kallal töötavad 3 Tartu Ülikooli Informaatika bakalaureuse 2. aasta tudengit. Probleemide korral on abiks aine Sissejuhatus andmeteadusesse juhendajad. Projektitöö põhineb kahel andmestikul, millest üks põhineb teisel, kuid on paremini vormistatud. Andmestikud sisaldavad suurt hulka andmeid Vormel 1 sarjast vahemikust 1950-2018 ehk spordi algusest kuni hetkeseisuga viimase lõpetatud hooajani. Riistvara külje pealt on projekti kallal töötamiseks erinevaid süle- ja lauaarvuteid. Tarkvarana kasutatakse programmeerimiskeelt Python 3 koos erinevate teekidega ja ka tarkvara Jupyter Notebook.

Nõudmised, oletused ja kitsaskohad

Projekti täieliku valmimise tähtaeg on 16. detsember 2019. Projekti esitamise tähtaeg posterisessioonil on 19. detsembril 2019, kell 14.00-17.00. Projektitööd võib lugeda arvestatavaks kui püstitatud eesmärkide kallal on oma võimete piires põhjalikult ja korralikult töötatud. Projekti lõpuks peab iga eesmärgiga seoses olema mingisuguseid tulemusi, mida posterisessioonil demonstreerida. Tulemuste olemasolu ei pea tähendama, et eesmärk täielikult õnnestus.

Riskid ja ettenägematud olukorrad

Projekti nõutav valmimise tähtaeg on lähedal sügissemestri lõpule, mis tähendab, et projekti kallal töötavatel tudengitel on palju tööd õppeainetega seoses, mis tähendab, et projekti läbiviimiseks võib jääda planeeritust vähem aega. Kuna on ka talveaeg ning tudengid käivad rahvarohketes kohtades, on ka haigestumise võimalus kõrgenenud.

Terminoloogia

Kõiki lahti seletamist vajavaid sõnu on hetkeseisuga raske määratleda, kuid siin on mõned neist.

- Hajuvusdiagramm - kahe tunnuse ühisjaotuse graafiline esitus, millelt saab välja lugeda erinevaid tendentse.
- Ennustusmudel - loogika (mudel), mis kasutab etteantud andmeid, et ennustada mingit tulemust.
- *Formula One Group* (lühendatult FOM) - korporatsioonide ühing, mis vastutab FIA Vormel 1 reklaamimise eest ning omab spordi kommertsõigusi.
- *Federation Internationale de l'Automobile* (lühendatult FIA) - rahvusvaheline ühing, mis esindab paljude võidusõiduseeriate huve.

Maksumus ja kasu

Kuna andmestikud on tasuta kättesaadavad ning kasutatavad, siis nende arvelt kulusi ei teki. Samuti tegelevad tudengid projekti kallal seoses ülikooli õppeainega, ei nõua projekti läbi viimine ka palkade maksmist. Riistvara projekti läbi viimiseks on eelnevalt olemas. Tarkvara projekti läbi viimiseks on tasuta allalaetav ja kasutatav. Kasu poole pealt saab *Formula One Group* kasvatada iga-aastast tulu (täpsetele numbritele pole võimalik ligi pääseda). Projekti läbiviijad saavad teadmisi andmekaeve ja andmete analüüsi poole pealt. Samuti saavad tudengid parandada oma teadmisi Vormel 1 sarja kohta ning arendada projekti- ja tiimitöö oskusi.

Andmekaeve eesmärkide seadmine

Andmekaeve eesmärgid

1. Luuakse 2014-2017 andmete põhjal efektiivse ennustusmudeli, mis suudab kvalifikatsioonide ja sõitude piiratud andmete põhjal ennustada tulevaste sõitude võitjaid (aastal 2018).
2. Luuakse graafikuid, mis illustreerivad raja kiireima ringi aja muutu aastatel 2004-2018. Vaadeldakse vaid radu, mida etteantud ajavahemikus kasutati tihedalt, ehk millel sõideti 11 või rohkem korda.
3. Luuakse 2012 aasta andmete põhjal graafikuid ja võrdlusi, mis illustreerivad ilmekalt viimase 10 aasta tasavägiseima hooaja kulgu.
4. Võrreldakse erinevate aastate andmete põhjal (pole ette määratletud) kahte edukaimat (maailmameistri tiitlite arvu poolest) sõitjat – Michael Schumacher ja Lewis Hamilton. Võrdluse tulemusena tuleb esitada aspektid, milles kumbki sõitja teisest tugevam oli.

Andmekaeve edukuse kriteeriumid

1. Ennustusmudeli loomine loetakse edukaks kui ta suudab 2018. aasta hooaja sõitude seas ennustada õigesti sõidu võitja vähemalt 33% sõitudest.
2. Graafikute loomine 2004-2018 aasta andmete põhjal loetakse edukaks, kui graafikult on selgelt ja ilmekalt näha radade kiireima ringi aja muutu etteantud vahemikus.
3. Graafikute ja võrdluste loomine 2012 aasta andmete põhjal loetakse edukaks, kui on välja toodud huvitavaid aspekte hooaja kulgemise kohta (huvitavuse määravad siinkohal projekti läbiviijad kriitilise pilguga).
4. Kahe Vormel 1 sarja edukaima sõitja võrdlus erinevate aastate andmete põhjal loetakse edukaks, kui võrdluse tulemusena tuuakse välja (ja visualiseeritakse) kummagi sõitja tugevaimad küljed.

Ülesanne 3

Andmete kogumine

Nõuded andmetele

Projekti edukaks läbiviimiseks on vaja pea kõiki avalikult kätte saadavaid andmeid Vormel 1 kohta. Täpsemalt oleks vaja andmeid Vormel 1 radade, tiimide, sõitjate, tiimide tulemuste, sõitjate tulemuste, ringiaegade, kvalifikatsiooni, sõitude, boksipeatuste, sõidu tulemuste ja

hooaegade kohta. Andmed võiks olla eeldatavasti .csv formaadis, mida on mugav projektitööks valitud keeles Python 3 töödelda. Andmete ajavahemik võiks olla alates Vormeli 1 algusest aastal 1950 kuni viimase lõpetatud hooajani aastal 2018.

Andmete saadavuse kontroll

Andmeid, mida projekti läbiviimiseks kasutatakse, on Internetist vabalt kättesaadavad. Kasutatavad andmed on .csv formaadis, seega on nende töötlemine mõnevõrra lihtsustatud. Andmete seast on suurima probleemina puudu kvalifikatsiooni ja sõitude ajal valitseva ilmastiku kohta. Samuti puuduvad sõitude kiireimad ringiajad enne 2004. aastat, mis oleks analüüsi läbiviimiseks kasulik. Lisaks kehtib üldine tendents, et mida varasema aastaga on tegemist, seda vähem on tollele aastale vastavat kättesaadavat informatsiooni. Sellepärast põhineb projekt pigem värskematel andmetel ning hetkeseisuga ei uurita süvitsi aastale 2004 eelnevaid andmeid. Ilmastiku andmetele mõistlikul viisil ligi pääseda pole võimalik, seega neid projektitöö käigus ei kasutata.

Kirjelda valiku kriteeriumeid

Andmete allikas: https://ergast.com/mrd/.

Kasutame valitud andmeallikat, kuna see on ainus (mida projekti läbiviijad leidsid) mahukas ja vabalt kasutatav andmebaas Vormel 1 kohta. Andmeallikas kasutame kõiki pakutud andmetabeleid, kuna need on omavahel tihedalt seotud. Pakutud andmetabelid on: circuits, constructor_results, constructor_standings, constructors, driver, driver_standings, lap_times, pit_stops, qualifying, races, results, seasons, status.

Andmetabelite siseselt kasutame suure tõenäosusega pea kõiki veerge (tunnuseid), kuid täpselt ei ole seda võimalik veel määratleda. Kuna andmetabeleid on rohkelt ning andmetabelites on tunnuseid palju, siis me ei hakka siinkohal kõiki valikuid ühe kaupa üles loetlema, vaid teeme seda andmestiku kirjelduse osana.

Andmestiku kirjeldus

Andmeallikast saame andmeid radade, tiimide, sõitjate, tiimide tulemuste, sõitjate tulemuste, ringiaegade, kvalifikatsiooni, sõitude, boksipeatuste, sõidu tulemuste ja hooaegade kohta.

Andmete allikas: https://ergast.com/mrd/. Andmetabelid on .csv formaadis. Internetis oleva andmestiku eelnev kirjeldus oli puudulik. Seega enne siinkohal olevat kirjeldust on andmestikule lisatud veerupealkirjad, mis mõtestatavad võimalikult täpselt, kuid lühidalt lahti veeru väljade sisu. Andmetabelid on üldjoontes vägagi põhjalikud ja sobivad projektitööks, kuid puudu on andmed ilmastiku kohta ning enne aastat 2004 on puudu ka kiireimate ringide ajad, mis oleks projekti läbiviimise jaoks kasulikud olnud.

Punktidega on allpool välja toodud andmetabelid. Iga andmetabli taga on sulgudes andmetabelis leiduvate ridade arv. Punktide alampunktidega on näidatud andmetabeli veerud (tunnused) ning nende lühikirjeldus.

- circuits (73 rida)
    - circuit_id - raja identifikaator numbrina
    - circuit_ref - raja identifikaator (lühend raja nimest) sõnena
    - name - raja täispikk nimi sõnena

    - location - raja asukoht (linn) sõnena
    - country - raja asukoht (riik) sõnena
    - lat - laiuskraad (koordinaadid) numbrina
    - lng - pikkuskraad (koordinaadid) numbrina
    - alt - kõrgus merepinnast numbrina
    - url - inglise keelse Wikipeedia link raja kohta sõnena
- constructor_standings (12305 rida)
    - constructor_standings_id - tiimi seisu identifikaator numbrina
    - race_id - sõidu identifikaator numbrina
    - constructor_id - identifikaator numbrina
    - points - punktide arv numbrina
    - position - positsioon numbrina
    - position_text - positsioon sõnena
    - wins - võitude arv numbrina
- constructor_results (11549 rida)
    - constructor_results_id - tiimi tulemuse identifikaator numbrina
    - race_id - sõidu identifikaator numbrina
    - constructor_id - tiimi identifikaator numbrina
    - points - punktide arv numbrina
    - status - tiimi tulemuse staatus (NULL kui konstuktor lõpetas hooaja tavapäraselt, muu kirjeldav väärtus, kui ei lõpetanud) sõnena
- constructors (208 rida)
    - constructor_id - tiimi identifikaator numbrina
    - constuctor_ref - tiimi identifikaator (lühend tiimi nimest) sõnena
    - name - tiimi täispikk nimi sõnena
    - nationality - tiimi asukohariik sõnena
    - url - inglise keelse Wikipedia link tiimi kohta sõnena
- drivers (846 rida)
    - driver_id - sõitja identifikaator numbrina
    - driver_ref - sõitja identifikaator (lühend sõitja nimest) sõnena
    - number - sõitja number numbrina
    - code - sõitja kolmetäheline lühend sõnena
    - forename - sõitja eesnimi sõnena
    - surname - sõitja perekonnanimi sõnena
    - birth_date - sõitja sünnikuupäev
    - nationality - sõitja kodakondsus sõnena
    - url - inglise keelse Wikipedia link sõitja kohta sõnena
- driver_standings (32545 rida)
    - driver_standings_id - sõitja seisu identifikaator numbrina
    - race_id - sõidu identifikaator numbrina
    - driver_id - sõitja identifikaator numbrina
    - points - punktide arv numbrina
    - position - sõitja positsioon üldarvestuses numbrina

    - position_text - sõitja positsioon üldarvestuses sõnena
    - wins - sõitja võitude arv hooajal numbrina

- lap_times (471428 rida)
    - race_id - sõidu identifikaator numbrina
    - driver_id - sõitja identifikaator numbrina
    - lap - ringi number numbrina
    - position - sõitja positsioon numbrina
    - time - ringi aeg mm:ss.ms formaadis sõnena
    - milliseconds - ringi aeg millisekundites numbrina
- pit_stops (7410 rida)
    - race_id - sõidu identifikaator numbrina
    - driver_id - sõitja identifikaator numbrima
    - stop - mitmes antud boksipeatus sama sõidu jooksul oli numbrina
    - lap - mitmendal ringil boksipeatus toimus numbrina
    - time - kellaaeg, mil boksipeatus toimus hh:mm:ss formaadis sõnena
    - duration - boksipeatuse kestus ss.ms formaadis sõnena
    - milliseconds - boksipeatuse kestus millisekundites numbrina
- qualifying (8333 rida)
    - qualify_id - kvalifikatsiooni identifikaator numbrina
    - race_id - sõidu identifikaator numbrina
    - driver_id - sõitja identifikaator numbrina
    - constructor_id - tiimi identifikaator numbrina
    - number - sõitja number numbrina
    - position - sõitja positsioon numbrina
    - q1 - esimese kvalifikatsiooniseeria kiireima ringi aeg mm:ss.ms formaadis sõnena
    - q2 - teise kvalifikatsiooniseeria kiireima ringi aeg mm:ss.ms formaadis sõnena
    - q3 - kolmanda kvalifikatsiooniseeria kiireima ringi aeg mm:ss.ms formaadis sõnena
- races (1039 rida)
    - race_id - sõidu indentifikaator numbrina
    - year - sõidu toimumisaasta numbrina
    - round - mitmenda selle aasta sõiduga oli tegu numbrina
    - circuit_id - raja identifikaator numbrina
    - name - sõidu ametlik nimi sõnena
    - date - sõidu toimumise kuupäev sõnena
    - time - sõidu alguse kellaaeg hh:mm:ss formaadis sõnena
    - url - inglise keelse Wikipedia link sõidu kohta sõnena
- results (24599 rida)
    - result_id - tulemuse identifikaator numbrina
    - race_id - sõidu identifikaator numbrina
    - driver_id - sõitja identifikaator numbrina

    - constructor_id - tiimi identifikaator numbrina
    - number - sõitja number numbrina
    - grid - stardipositsioon stardiruudustikul numbrina
    - position - lõpetamispositsioon numbrina
    - position_text - lõpetamispositsiooni number sõnena või katkestamise tähis sõnena
    - position_order - lõpetamispositsioon kui kaasa arvata katkestamised sõnena
    - points - saadud punktide arv numbrina
    - laps - sõidetud ringide arv numbrina
    - time - sõidu kogukestvus hh:mm:ss formaadis sõnena
    - milliseconds - sõidu kogukestvus millisekundites numbrina
    - fastest_lap - ring, millel tehti kiireim ringi aeg numbrina
    - rank - sõidu kiireimate ringide edetabeli positsiooni numbrina
    - fastest_lap_time - kiireima ringi aeg mm:ss.ms formaadis sõnena
    - fastest_lap_speed - kiireim kiirus numbrina
    - status_id - sõidu lõpetamise staatus (lõpetas tavapäraselt, katkestas käigukasti pärast jne.) sõnena
- seasons (70 rida)
    - year - hooaja aastaarv numbrina
    - url - inglise keelse Wikipedia link hooaja kohta
- status (134 rida)
    - status_id - staatuse identifikaator numbrina
    - status - staatuse kirjeldus (lõpetas tavapäraselt, katkestas käigukasti pärast jne.) sõnena

Andmete uurimine

Kuna andmetabelites on väga suures koguses andmeid, ei olnud siinkohal mõtekas täpsustada iga tunnuse miinimum- ja maksimumväärtusi. Siiani läbi viidud vaatluste ja katsetuste käigus (suuri) anomaaliaid leidunud pole. Suurima probleemina on andmestikust puudu andmed kvalifikatsiooni ja sõitude ajal valitsevate ilmastikuolude kohta. Kuigi selle olemasolu andmestiku looja otseselt lubanud pole, oleks see vägagi kasulik tunnus, mida eri ülesannete käigus kasutada. Andmestikus eksisteerivad sõitude kiireimate ringide ajad vaid alates aastast 2004, mis muudab keerukamaks ühe eesmärgi – spordi edukaimate sõitjate Lewis Hamiltoni ja Michael Schumacheri võrdluse – läbiviimist. Üldjoontes, mida varasema aastaga on tegemist, seda puudulikumad on andmestikus olevad andmed.

Miinimum- ja maksimumväärtused andmetabelites:

- circuits
  - circuit_id - min: 1, max: 74
  - circuit_ref - min: BAK, max: zolder
  - name - min: A1-Ring, max: Zolder
  - location - min: Abu Dhabi, max: Zandvoort
  - country - min: Argentina, max: Vietnam
  - lat - min: -37.8497, max: 57.2653
  - lng - min: -118.189, max: 144.968
  - alt - min: puudub (min() käsk ei andnud tulemust), max: puudub (max() käsk ei andnud tulemust
  - url - min: http://en.wikipedia.org/wiki/A1-Ring, max: http://en.wikipedia.org/wiki/Zolder
- constructor_standings
  - constructor_standings_id - min: 1.0, max: 27452
  - race_id - min: 1.0, max: 1029
  - constructor_id - min: 1.0, max: 211
  - points - min: 0.0, max: 765
  - position - min: 1.0, max: 22
  - position_text - min: 1.0, max: E
  - wins - min: 0.0, max: 19
- constructor_results
  - constructor_results_id - min: 1.0, max: 16049.0
  - race_id - min: 1.0, max: 1029.0
  - constructor_id - min: 1.0, max: 211.0
  - points - min: 0.0, max: 66.0
  - status - min: puudub (min() käsk ei andnud tulemust), max: puudub (max() käsk ei andnud tulemust
- constructors
  - constructor_id - min: 1, max: 211
  - constuctor_ref - min: adams, max: zakspeed
  - name - min: AFM, max: Zakspeed
  - nationality - min: American, max: Swiss
  - url - min: http://en.wikipedia.org/wiki/A.J._Watson, max: http://en.wikipedia.org/wiki/Zakspeed
- drivers
  - driver_id - min: 1, max: 848
  - driver_ref - min: Cannoc, max: zunino
  - number - min: puudub (min() käsk ei andnud tulemust), max: puudub (max() käsk ei andnud tulemust
  - code - min: puudub (min() käsk ei andnud tulemust), max: puudub (max() käsk ei andnud tulemust
  - forename - min: Adolf, max: Óscar

    - surname - min: Abate, max: Étancelin
    - birth_date - min: 1896-12-28, max: 1999-11-13
    - nationality - min: American, max: Venezuelan
    - url - min: puudub (min() käsk ei andnud tulemust), max: puudub (max() käsk ei andnud tulemust
- driver_standings
    - driver_standings_id - min: 1.0, max: 69728
    - race_id - min: 1.0, max: 1029
    - driver_id - min: 1.0, max: 848
    - points - min: 0.0, max: 408
    - position - min: 1.0, max: 108
    - position_text - min: 1.0, max: D
    - wins - min: 0.0, max: 13
- lap_times
    - race_id - min: 1, max: 1029
    - driver_id - min: 1, max: 848
    - lap - min: 1, max: 78
    - position - min: 1, max: 24
    - time - min: 10:32.179, max: 9:45.712
    - milliseconds - min: 66957, max: 7507547
- pit_stops
    - race_id - min: 841, max: 1029
    - driver_id - min: 1, max: 848
    - stop - min: 1, max: 6
    - lap - min: 1, max: 74
    - time - min: 13:04:31, max: 22:04:12
    - duration - min: 12.897, max: 59.555
    - milliseconds - min: 12897, max: 2011266
- qualifying
    - qualify_id - min: 1, max: 8357
    - race_id - min: 1, max: 1029
    - driver_id - min: 1, max: 848
    - constructor_id - min: 1, max: 211
    - number - min: 0, max: 99
    - position - min: 1, max: 28
    - q1 - min: puudub (min() käsk ei andnud tulemust), max: puudub (max() käsk ei andnud tulemust
    - q2 - min: puudub (min() käsk ei andnud tulemust), max: puudub (max() käsk ei andnud tulemust
    - q3 - min: puudub (min() käsk ei andnud tulemust), max: puudub (max() käsk ei andnud tulemust

- races
    - race_id - min: 1, max: 1052
    - year - min: 1950, max: 2020
    - round - min: 1, max: 22
    - circuit_id - min: 1 , max: 74
    - name - min: Abu Dhabi Grand Prix, max: Vietnamese Grand Prix
    - date - min: 1950-05-13, max: 2020-11-29
    - time - min: puudub (min() käsk ei andnud tulemust), max: puudub (max() käsk ei andnud tulemust
    - url - min: http://en.wikipedia.org/wiki/1950_Belgian_Grand_Prix, max: https://en.wikipedia.org/wiki/2020_Vietnamese_Grand_Prix
- results
    - result_id - min: 1, max: 24605
    - race_id - min: 1, max: 1029
    - driver_id - min: 1, max: 848
    - constructor_id - min: 1, max: 211
    - number - min: puudub (min() käsk ei andnud tulemust), max: puudub (max() käsk ei andnud tulemust
    - grid - min: 0, max: 34
    - position - min: puudub (min() käsk ei andnud tulemust), max: puudub (max() käsk ei andnud tulemust
    - position_text - min: 1, max: W
    - position_order - min: 1, max: 39
    - points - min: 0, max: 50
    - laps - min: 0, max: 200
    - time - min: puudub (min() käsk ei andnud tulemust), max: puudub (max() käsk ei andnud tulemust
    - milliseconds - min: puudub (min() käsk ei andnud tulemust), max: puudub (max() käsk ei andnud tulemust
    - fastest_lap - min: puudub (min() käsk ei andnud tulemust), max: puudub (max() käsk ei andnud tulemust
    - rank - min: puudub (min() käsk ei andnud tulemust), max: puudub (max() käsk ei andnud tulemust
    - fastest_lap_time - min: puudub (min() käsk ei andnud tulemust), max: puudub (max() käsk ei andnud tulemust
    - fastest_lap_speed - min: puudub (min() käsk ei andnud tulemust), max: puudub (max() käsk ei andnud tulemust
    - status_id - min: 1, max: 137
- seasons
    - year - min: 1950, max: 2020
    - url - min: https://en.wikipedia.org/wiki/1950_Formula_One_season, max: https://en.wikipedia.org/wiki/2020_Formula_One_World_Championship

- status
    - status_id - min: 1, max: 137
    - status - min: +1 Lap, max: Withdrew

Põjalikum väärtuste analüüs on andmetabelite ja neis leiduvate tunnuste rohkuse tõttu hetkese projekti kirjelduse jaoks liiga mahukas. Seega on siinkohal välja toodud vaid tunnuste miinimum- ja maksimumväärtused andmetabelites. Väärib mainimist, et andmetabelites on mitmeid sõne kujul väärtusi, mis tegelikult kujutavad nt kuupäeva või aega mingis keerulisemas formaadis. Samuti on veerge (tunnuseid), kus on segamini arve ja tähti, mis on tegelikkuses seal loogiliselt, kuid nt *DataFrame* käskudele min() ja max() segased. Seega siinkohal välja toodud väärtused ei pruugi olla otseselt loogilised või mõistetavad. Ülal enne miinimum- ja maksimumväärtuste välja toomist on mainitud ka andmetabelites puuduvaid andmeid, mis oleks projekti töö käigus kasulikuks osutunud, kuid mida ei õnnestunud muudest allikatest leida.

Andmete kvaliteedikontroll

Nagu varasemalt projekti kirjelduses välja toodud, on andmestikust puudu kiireimad ringiajad enne 2004. aastat. Samuti on täielikult puudu ilmastikuandmed, kuigi ilmastik mängib võidusõidus üpriski suurt rolli. Lisaks, mida varasemaid sõite vaadelda, seda vähem on nende kohta andmeid või on andmed ebatäpsemad. Seetõttu keskendutakse projekti eesmärkides võidusõitudele, mis toimusid aastal 2004 või hiljem. Varasemaid andmeid eriti põhjalikult ei uurita. Ilmastiku roll jääb siinkohal tulemustest kõrvale, kuna asendusandmeid ilmastiku kohta ei suudetud leida. Olemasolevate ja värskemate (2004 või hiljem) andmete seas pole hetkeseisuga kvaliteediprobleeme tuvastatud ning need sobivad hästi eesmärkide läbiviimiseks.

<u>Ülesanne 4</u>

Projektiplaan koos ülesannetega

- Puhastatakse ja kujundatakse algseid andmeid sobivaks. **(5 tundi, 2.5h Aveli Klaos ja Siim Tanel Laisaar mõlemad)**
- Luuakse 2014-2018 aasta andmete põhjal ennustusmudel, mis ennustab kvalifikatsioonide ja sõitude piiratud andmete põhjal sõitude võitjaid (esialgu aastal 2018) **(25 tundi, 25h Siim Tanel Laisaar)**
    - Valmistatakse kogu andmestiku põhjal andmetabel, mis sisaldab sobivaid andmeid 2014-2018 aastatest. Filtreeritakse seejärel andmetabelit nii, et see sisaldaks andmeid vaid nende 6 sõitja kohta, kes aastal 2018 sõitsid Mercedese, Red Bulli või Ferrari tiimis (sh andmed, mis tekkisid sõitjate kohta varasemates tiimides antud ajavahemikus).
    - Teisendatatakse ebasobivas formaadis veerge (tunnuseid).
    - Asendatatakse puuduvad väljad ülejäänud 5 sõitja keskmistega.
    - Lisatakse 2017 ja 2018 aasta andmetele juurde sõidu võitjad.
    - Eraldatakse 2014-2016 aastast ette valmistatud andmed *training set*-iks, 2017 aasta andmed *validation set*-iks ja 2018 aasta andmed *test set*-iks.
    - Katsetatakse esmalt ennustamist *Random Forest Classifier*iga.

    - Optimeeritakse ja parendatatakse valitud *classifier*-it ning katsetatakse ka teisi, et leida optimaalseim.
    - Visualiseeritakse ennustusmudeli efektiivsust.
    - Lisatakse ennustusmudeli tööd ja efektiivsust lahti seletav lühikirjeldus.

- Luuakse graafikuid, mis illustreerivad raja kiireima ringi aegu ajavahemikus 2004-2018. Vaadeldakse radu, millel on etteantud ajavahemikus sõidetud 11 või rohkem korda. **(10-15 tundi, 2.5h Siim Tanel Laisaar, 7.5h+ Aveli Klaos)**
    - Valmistatakse kogu andmestiku põhjal andmetabel, mis sisaldab sobivaid andmeid 2004-2018 aastatest.
    - Teisendatakse ebasobivas formaadis veerge (tunnuseid).
    - Filtreeritakse andmetabelit nii, et see sisaldaks andmeid vaid radade kohta, millel on etteantud ajavahemikus sõidetud 11 või rohkem korda.
    - Moodustatakse iga vaadeldava raja kohta hajuvusdiagramm. Diagrammi x-teljel on aastad 2004-2018 ning y-teljel raja kiireima ringi ajad sekundites.
    - Kujundatakse loodud graafikuid visuaalselt selgeteks ja huvitatavateks.
    - Lisatakse graafikuid lahti seletav lühikirjeldus.
- Luuakse graafikuid, mis iseloomustavad 2012. aasta kui viimase dekaadi ühe tasavägiseima aasta hooaja kulgu. **(25 tundi, 15h Aveli Klaos, 5h Peeter Paal ja Siim Tanel Laisaar mõlemad)**
    - Valmistatakse kogu andmestiku põhjal andmetabeleid, mis sisaldavad sobivaid andmeid 2012. aastast.
    - Leitakse ja filtreeritakse andmetabeleid nii, et alles jääksid veerud (tunnused), mida on huvitav võrrelda.
    - Moodustatakse võrreldavate tunnuste põhjal graafikuid.
    - Kujundatakse loodud graafikuid visuaalselt selgeteks ja huvitavateks.
    - Lisatakse graafikuid lahti seletav lühikirjeldus.
- Luuakse graafikuid ja võrreldakse kahte Vormel 1 edukaimat (maailmameistri tiitlite arvu poolest) sõitjat – Michael Schumacherit ja Lewis Hamiltoni. Tuuakse välja aspektid, milles kumbki sõitja teisest tugevam oli. **(25 tundi, 25h Peeter Paal)**
    - Valmistatakse kogu andmestiku põhjal andmetabeleid, mis sisaldavad sobivaid andmeid erinevatest aastatest.
    - Leitakse ja filtreeritakse andmetabeleid nii, et alles jääksid veerud (tunnused), mida on huvitav võrrelda.
    - Moodustatakse võrreldavate tunnuste põhjal graafikuid.
    - Kujundatakse loodud graafikuid visuaalselt selgeteks ja huvitavateks.
    - Valitakse välja spetsiifilised võrreldavad aspektid, milles kumbki sõitja teisest tugevam oli. Võrdlust toetatakse vajadusel lüihkese tekstiga.
    - Lisatakse graafikuid ja võrdlusi lahti seletav lühikirjeldus.

- Projekti tulemused kujundatakse nii, et neid oleks võimalik demonstreerida posteri formaadis. **(5-10 tundi, 2.5h+ Siim Tanel Laisaar, Peeter Paal, Aveli Klaos kõik)**
    - Püstitatud eesmärkide tulemused kujundatakse posteril esitlemiseks sobivaks.
    - Eesmärkide tulemused kujundatakse ühiseks tervikuks.

Tundide jaotus ülesannete juures pole rangelt määratletud ja võib muutuda olenevalt sellest, mis hetkel kellelgi rohkem aega on ja ka sellest, kui keegi jääb oma ülesandega hätta ja vajab tiimikaaslastelt abi.

Meetodid ning vahendid, mida plaanime kasutada

Projekti läbiviimisel on peamiseks programmeerimiskeeleks Python 3, mida kasutatakse Jupyter Notebook tööriista siseselt. Python 3-ga ühiselt kasutatakse sellel olemasolevaid teeke (nt pandas, numpy, matplotlib, seaborn jt). Riistvaraks on olemasolevad süle- ja lauaarvutid.