Lokaverkefni

Fríða Margrét Guðmundsdóttir

16.4.2025

Contents

Inngangur

Markmið þessa lokaverkefnis er að beita tölfræðigreiningu til að skoða tengsl og áhrif milli breyta í gagnasafni sem inniheldur tæknilegar og lýsandi upplýsingar um bíla. Gagnasafnið inniheldur bæði flokkabreytur (eins og eldsneytistegund og gerð bíls) og samfelldar talnabreytur (líkt og hestöfl og eldsneytiseyðslu).

Gögnin voru sótt af Kaggle, nánar tiltekið úr gagnasafninu *Car Features and MSRP* sem inniheldur yfir 1000 skráningar með upplýsingum um bílategundir, eldsneyti, afköst og verð. Gagnasafnið var lítillega aðlagað fyrir greiningu með því að velja aðeins fjórar algengustu eldsneytistegundirnar til þess að bæta skýrleika og einfaldleika framsetningarinnar.

Í verkefninu eru settar fram þrjár rannsóknarspurningar sem svarað er með viðeigandi tölfræðilegum aðferðum. Fyrsti hlutinn skoðar tengsl tveggja flokkabreyta með kí-kvaðrat prófi. Annar hlutinn ber saman meðaltöl samfelldrar breytu milli hópa með ANOVA og Kruskal–Wallis prófum. Þriðji hlutinn metur línulegt samband milli tveggja samfelldra breyta með Pearson fylgniprófi.

Greiningin byggir á sjónrænum framsetningum, lýsistærðum og tilgátuprófum. Áhersla er lögð á túlkun niðurstaðna, mat á forsendum og umræðu um mögulega bjaga sem geta haft áhrif á ályktanir. Verkefnið sýnir hvernig tölfræði getur dregið fram gagnleg mynstur úr flóknum og fjölbreyttum gögnum.

```
library(knitr)
library(kableExtra)

tolfraedi_tafla <- data.frame(
  "Liður" = c("1", "2", "3"),
  "Breytur" = c(
    "Vehicle Style, Engine Fuel Type",
    "Engine Fuel Type, City MPG",
    "Engine HP, City MPG"
  ),
  "Rannsóknarspurning" = c(
```

```
    "Er samband milli gerðar bíls og eldsneytistegundar?",
    "Er munur á borgaraksturseyðslu eftir eldsneytistegund?",
    "Er fylgni milli hestafla og sparneytni í borg?"
  ),
  "Tölfræðipróf" = c(
    "Kí-kvaðrat próf",
    "ANOVA og Kruskal-Wallis",
    "Pearson fylgnipróf"
  )
)

kable(tolfraedi_tafla, format = "latex", align = "c",
      caption = "Yfirlit yfir rannsóknarspurningar, breytur og próf", booktabs = TRUE) %>%
  kable_styling(latex_options = c("striped", "hold_position")) %>%
  column_spec(1, width = "1cm") %>%
  column_spec(2, width = "4cm") %>%
  column_spec(3, width = "5cm") %>%
  column_spec(4, width = "3.5cm")
```

Table 1: Yfirlit yfir rannsóknarspurningar, breytur og próf

Liður	Breytur	Rannsóknarspurning	Tölfræðipróf
1	Vehicle Style, Engine Fuel Type	Er samband milli gerðar bíls og eldsneytistegundar?	Kí-kvaðrat próf
2	Engine Fuel Type, City MPG	Er munur á borgaraksturseyðslu eftir eldsneytistegund?	ANOVA og Kruskal–Wallis
3	Engine HP, City MPG	Er fylgni milli hestafla og sparneytni í borg?	Pearson fylgnipróf

Liður 1 - Tvær flokkabreytur

Er samband milli gerðar bíls og eldsneytistegundar?

Við viljum kanna hvort samband sé á milli þess hvernig bíll er hannaður (t.d. sem fólksbíll, jeppi eða sendibíll) og hvers konar eldsneyti hann notar. Til þess skoðum við sambandið milli flokkabreyta: *Vehicle Style* (gerð bíls) og *Engine Fuel Type* (eldsneytistegund).

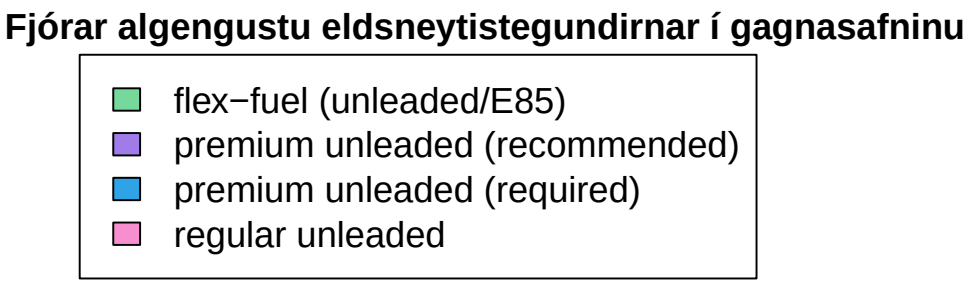
Við byrjum á að teikna súlurit sem sýnir hlutfall mismunandi eldsneytistegunda innan hverrar bílagerðar í gagnasafninu. Hlutföllin eru reiknuð með prop.table(..., margin = 1) sem þýðir að þau eru miðuð við heildarfjölda bíla í hverri bílagerð, þ.e. hlutfall hvers eldsneytis reiknað út frá fjölda bíla innan viðkomandi hóps en ekki út frá heildarfjölda bíla í gagnasafninu.

Til að einfalda framsetninguna og bæta læsileika grafsins ákváðum við að einblína á fjórar algengustu eldsneytistegundirnar í gagnasafninu. Þessar fjórar tegundir mynda stærstan hluta allra skráninga en sjaldgæfari eldsneytistegundir voru með mjög fá dæmi og myndu bæta lítið við greininguna.

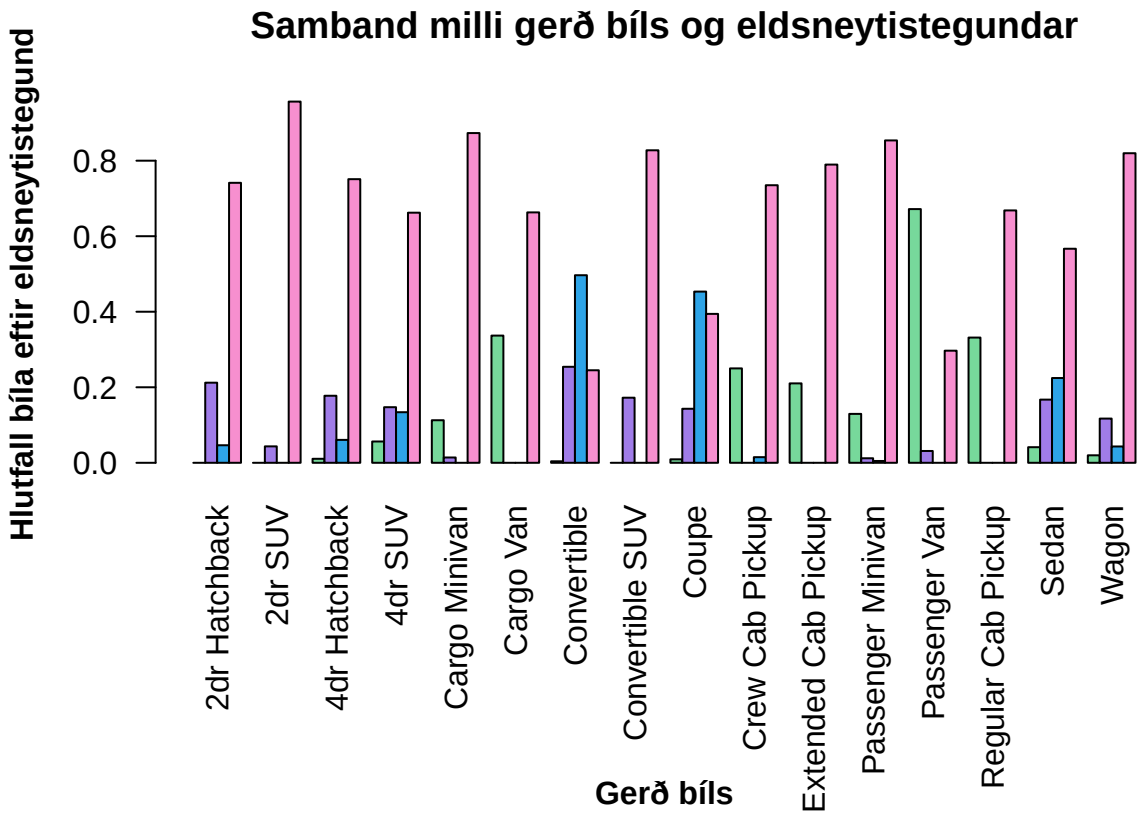
```
unzip("data.csv.zip")
data <- read.csv("data.csv", stringsAsFactors = TRUE)

algengasta_eldsneytid <- names(sort(table(data$Engine.Fuel.Type), decreasing = TRUE))[1:4]
data2 <- subset(data, Engine.Fuel.Type %in% algengasta_eldsneytid)
```

```r
data2$Engine.Fuel.Type <- droplevels(data2$Engine.Fuel.Type)

tafla <- table(data2$Vehicle.Style, data2$Engine.Fuel.Type)
prop_tafla <- prop.table(tafla, margin = 1)

litir  <- c("#77d89b", "#a07ce8",  "#2ea4e7", "#f78fcf")
```

```r
plot.new()

text(x = 0.48, y = 0.95,
     labels = "Fjórar algengustu eldsneytistegundirnar í gagnasafninu",
     font = 2, cex = 1.2,)

legend("topright",
       x = 0.05, y = 0.85,
       legend = colnames(prop_tafla),
       fill = litir,
       cex = 1.2)
```




```r
par(mar = c(9, 4, 4, 2))  # (bottom, left, top, right)

barplot(t(prop_tafla),
        beside = TRUE,
        col = litir,
        las = 2,
        xlab = "",
        ylab = "",
        main = "Samband milli gerð bíls og eldsneytistegundar")

mtext("Gerð bíls", side = 1, line = 8, font = 2)

mtext("Hlutfall bíla eftir eldsneytistegund", side = 2, line = 3, font = 2)
```



Í grafinu hér að ofan sjáum við hvernig dreifing eldsneytistegunda breytist eftir gerð bíls. Til dæmis má sjá að eldsneytið „regular unleaded“ er ríkjandi hjá flestum bílastílum en „premium unleaded“ (bæði recommended og required) birtist í meiri mæli hjá bílum eins og Coupe, Convertible og sumum Pickup gerðum.

Þetta bendir til þess að sportlegir eða kraftmeiri bílar séu frekar hannaðir til að nota sérstakt eldsneyti eins og „premium unleaded“ sem hentar öflugum vélum betur. Á hinn bóginn má sjá að bílar eins og Hatchback og Sedan nota fjölbreyttari eldsneytistegundir á meðan gerðir eins og Passenger Van eru með meiri einsleitni í eldsneytisvali.

Súlurnar sýna hlutfallslega dreifingu eldsneytistegunda innan hverrar bílagerðar en ekki heildarfjölda bíla í hverjum flokki. Því sést ekki beint í grafinu hversu margir bílar tilheyra hverri gerð sem þýðir að stærðir hópanna eru ekki sýnilegar og þarf að hafa það í huga við túlkun niðurstaðna.

Tilgátupróf

Til að kanna hvort samband sé á milli gerðar bíls (*Vehicle Style*) og eldsneytistegundar (*Engine Fuel Type*) má nota kí-kvaðrat próf. Prófið ber saman rauntíðni og væntitíðni í krosstöflu og metur hvort dreifingin sem sést í gögnunum sé líkleg til að vera tilviljanakennd. Ef p-gildið er < 0.05 teljum við samband vera milli breytanna.

```
kikvadrat <- chisq.test(tafla)
```

```
## Warning in chisq.test(tafla): Chi-squared approximation may be incorrect
```

```r
library(knitr)
library(kableExtra)

prof_nidurstada <- data.frame(
  "Próf" = "Kí-kvaðrat",
  "X²" = round(kikvadrat$statistic, 2),
  "Frígráður" = kikvadrat$parameter,
  "P-gildi" = format.pval(kikvadrat$p.value, digits = 3, eps = .001)
)

kable(prof_nidurstada, format = "latex", align = "c", caption = "Niðurstaða prófsins", booktabs = TRUE) %>%
  kable_styling(latex_options = c("striped", "hold_position")) %>%
  column_spec(1:4, width = "3cm")
```

Table 2: Niðurstaða prófsins

	Próf	X.	Frígráður	P.gildi
X-squared	Kí-kvaðrat	4314.71	45	<0.001

P-gildið er mjög lágt (< 0.001) sem þýðir að við höfnum núlltilgátunni með sterkum rökum. Það bendir til þess að marktækt samband sé milli gerðar bíls og eldsneytistegundar í þessu gagnasafni. Með öðrum orðum, dreifing eldsneytis fer eftir því hvaða bílastíl um ræðir og þessar tvær flokkabreytur eru ekki óháðar.

Viðvörunin „Chi-squared approximation may be incorrect“ bendir til þess að einhver væntigildi í krosstöflunni séu mjög lítil og því gætu forsendur kí-kvaðrat prófsins verið brotnar að einhverju leyti.

Þrátt fyrir að sum væntigildi í kí-kvaðrat prófinu hafi verið tiltölulega lág þá var taflan nægilega stór og dreifing gildanna slík að áhrif þess á marktækni niðurstöðunnar eru talin óveruleg. Því teljum við að túlkun prófsins haldi.

Bjagar sem geta skekkt niðurstöðurnar

Eldsneytistegundir valdar:

Við völdum aðeins fjórar algengustu eldsneytistegundirnar til að bæta læsileika grafsins. Þetta einfaldaði greininguna en útilokar sjaldgæfari valkosti og getur rýrt dýpt niðurstöðunnar.

Túlkun tengsla:

Kí-kvaðrat prófið sýnir aðeins hvort tengsl séu milli breyta, ekki hvort önnur orsaki hina. Því má ekki álykta að bílastíll valdi vali á eldsneyti, aðeins að þau tengjast í þessum gögnum.

Úrtaksbjagi:

Ekki liggur fyrir hvernig gagnasafnið var tekið saman. Ef það byggir eingöngu á tilteknum tegundum bíla (t.d. eingöngu amerískir bílar) eða tilteknum framleiðendum gæti úrtakið verið ófullnægjandi fyrir alla bíla. Slíkur úrtaksbjagi getur skekkt niðurstöðurnar.

Samantekt

Grafið hér að ofan bendir til þess að sportlegir eða kraftmeiri bílar séu frekar hannaðir til að nota sérstakt eldsneyti eins og „premium unleaded“. Á hinn bóginn má sjá að bílategundir eins og Hatchback og Sedan nota fjölbreyttari eldsneytistegundir en gerðir eins og Passenger Van sýna meiri einsleitni í eldsneytisvali.

Samantekið benda niðurstöður til þess að marktækt samband sé milli gerðar bíls og eldsneytistegundar í gagnasafninu. Hins vegar ber að hafa í huga takmarkanir greiningarinnar svo sem möguleika á úrtaksbjaga og einföldun með val á eldsneytistegundum. Þrátt fyrir það veita niðurstöðurnar áhugaverða og gagnlega innsýn í hvernig þessar tvær breytur tengjast. Þær geta endurspeglað hvernig framleiðendur hanna bíla með mismunandi notkun í huga þar sem öflugari ökutæki eru hönnuð fyrir afköst og hraða en aðrir bílar fyrir daglega notkun og hagkvæmni. Gögnin geta því nýst bæði neytendum sem vilja taka upplýsta ákvörðun um rekstrarkostnað og framleiðendum sem vilja staðfæra hönnun eftir þörfum mismunandi markhópa.

Liður 2 - Talnabreyta og flokkabreyta

Er munur á meðaleyðslu í borg eftir eldsneytistegund?

Við viljum kanna hvort meðaleyðsla bíla í borgarakstri sé mismunandi eftir því hvaða eldsneyti þeir nota. Við skoðum því samband milli flokkabreytu, *Engine Fuel Type* (eldsneytistegund) og talnabreytu, *city.mpg* (eyðsla í borg).

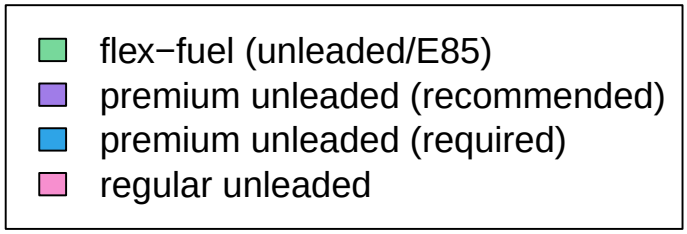
Við byrjum á því að teikna boxplot sem sýnir dreifingu eyðslunnar innan hverrar eldsneytistegundar. Þetta gefur okkur innsýn í meðaleyðslu og breytileika innan hópa.

Til að einfalda framsetninguna og bæta læsileika grafsins ákváðum við að einblína á fjórar algengustu eldsneytistegundirnar í gagnasafninu eins og gert var í lið 1. Þessar fjórar tegundir mynda stærstan hluta allra skráninga en sjaldgæfari eldsneytistegundir voru með mjög fá dæmi og myndu bæta lítið við greininguna.

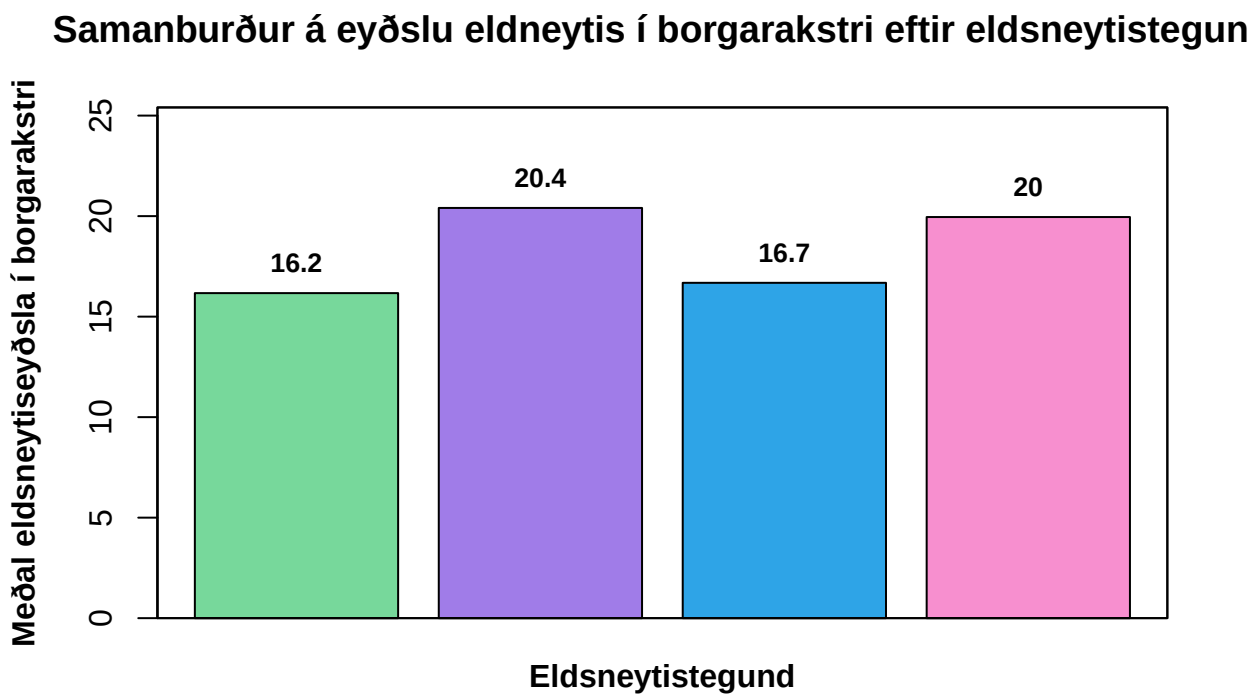
```
data_lidur2 <- subset(data, Engine.Fuel.Type %in% algengasta_eldsneytid & !is.na(city.mpg))
data_lidur2$Engine.Fuel.Type <- droplevels(data_lidur2$Engine.Fuel.Type)
medaltol <- tapply(data_lidur2$city.mpg, data_lidur2$Engine.Fuel.Type, mean)
```

```
plot.new()

text(x = 0.48, y = 0.95,
     labels = "Fjórar algengustu eldsneytistegundirnar í gagnasafninu",
     font = 2, cex = 1.2,)

legend("topright",
       x = 0.05, y = 0.85,
       legend = colnames(prop_tafla),
       fill = litir,
       cex = 1.2)
```

Fjórar algengustu eldsneytistegundirnar í gagnasafninu




```r
medaltol <- tapply(data_lidur2$city.mpg, data_lidur2$Engine.Fuel.Type, mean)

stadsetningar <- barplot(medaltol,
                         col = litir,
                         ylim = c(0, max(medaltol) + 5),
                         xlab = "",
                         ylab = "",
                         main = "Samanburður á eyðslu eldneytis í borgarakstri eftir eldsneytistegund",
                         names.arg = FALSE, )

box()
mtext("Eldsneytistegund", side = 1, line = 1, font = 2)

mtext("Meðal eldsneytiseyðsla í borgarakstri", side = 2, line = 3, font = 2)

text(x = stadsetningar,
     y = medaltol + 1.5,
     labels = round(medaltol, 1),
     cex = 0.8,
     font = 2)

box()
```



Stöplaritið sýnir meðaleldsneytiseyðslu í borgarakstri eftir fjórum algengustu eldsneytistegundunum í gagnasafninu. Bílar sem nota "premium unleaded (recommended)" skera sig úr með hæstu meðaleyðslu, 20.4 MPG (miles per gallon), á meðan bílar sem nota "flex-fuel (unleaded/E85)" sýna lægstu meðaleyðsluna,16.2 MPG. Aðrar eldsneytistegundir eins og "regular unleaded" og "premium unleaded (required)" eru nær meðaltali með 20.0 og 16.7 MPG.

Þessar niðurstöður gefa til kynna að bílar hannaðir fyrir "premium recommended" eldsneyti séu hugsanlega nýrri, minni eða sérhannaðir fyrir meiri sparneytni. Á hinn bóginn eru bílar sem nota "flex-fuel" eldsneyti stærri eða öflugri ökutæki sem gæti skýrt hærri eldsneytiseyðslu í borg.

Grafið undirstrikar þannig mikilvægi þess að skoða eldsneytisval í samhengi við hönnun og afköst ökutækja þegar verið er að greina sparneytni. Því má ályktа að eldsneytistegund sé tengd við meðaleyðslu og hafi raunveruleg áhrif á sparneytni í borgarakstri.

Tilgátupróf

ANOVA-próf

Til að kanna hvort meðaleyðsla bíla í borgarakstri sé mismunandi eftir eldsneytistegund (*Engine Fuel Type*) má nota einfalt einhliða ANOVA-próf.

Prófið metur hvort munur á meðaltölum sem sést í gögnunum sé líklegur til að stafa af tilviljun. Ef p-gildið er < 0.05 teljum við að marktækur munur sé á meðaleyðslu milli eldsneytistegunda, þ.e. að minnsta kosti einn hópur sé frábrugðinn hinum.

```
anova_prof <- aov(city.mpg ~ Engine.Fuel.Type, data = data_lidur2)
```

```r
library(knitr)
library(kableExtra)

anova_summary <- summary(anova_prof)[[1]]

anova_tafla <- data.frame(
  "Þáttur" = rownames(anova_summary),
  "Frígráður (Df)" = anova_summary[["Df"]],
  "Sum Sq" = round(anova_summary[["Sum Sq"]], 1),
  "Mean Sq" = round(anova_summary[["Mean Sq"]], 1),
  "F-gildi" = round(anova_summary[["F value"]], 1),
  "P-gildi" = format.pval(anova_summary[["Pr(>F)"]], digits = 3, eps = 0.001)
)

kable(anova_tafla, format = "latex", caption = "Niðurstaða ANOVA-prófs", booktabs = TRUE) %>%
  kable_styling(latex_options = c("striped", "hold_position")) %>%
  column_spec(1:6, width = "2.5cm")
```

Table 3: Niðurstaða ANOVA-prófs

Þáttur	Frígráður..Df.	Sum.Sq	Mean.Sq	F.gildi	P.gildi
Engine.Fuel.Type	3	27275.9	9092.0	322.4	<0.001
Residuals	11599	327053.0	28.2	NA	NA

Taflan sýnir niðurstöður ANOVA-prófs sem bar saman meðaleyðslu (city MPG) eftir fjórum eldsneytistegundum.

Frígráður (Df) eru þrjár fyrir hópana (fjöldi hópa - 1) og 11.599 fyrir villuna (heildarfjöldi - fjöldi hópa).

F-gildið er 322 og p-gildið < 0.001 sem gefur til kynna mjög marktækan mun á meðaleyðslu milli a.m.k. einnar eldsneytistegundar og hinna.

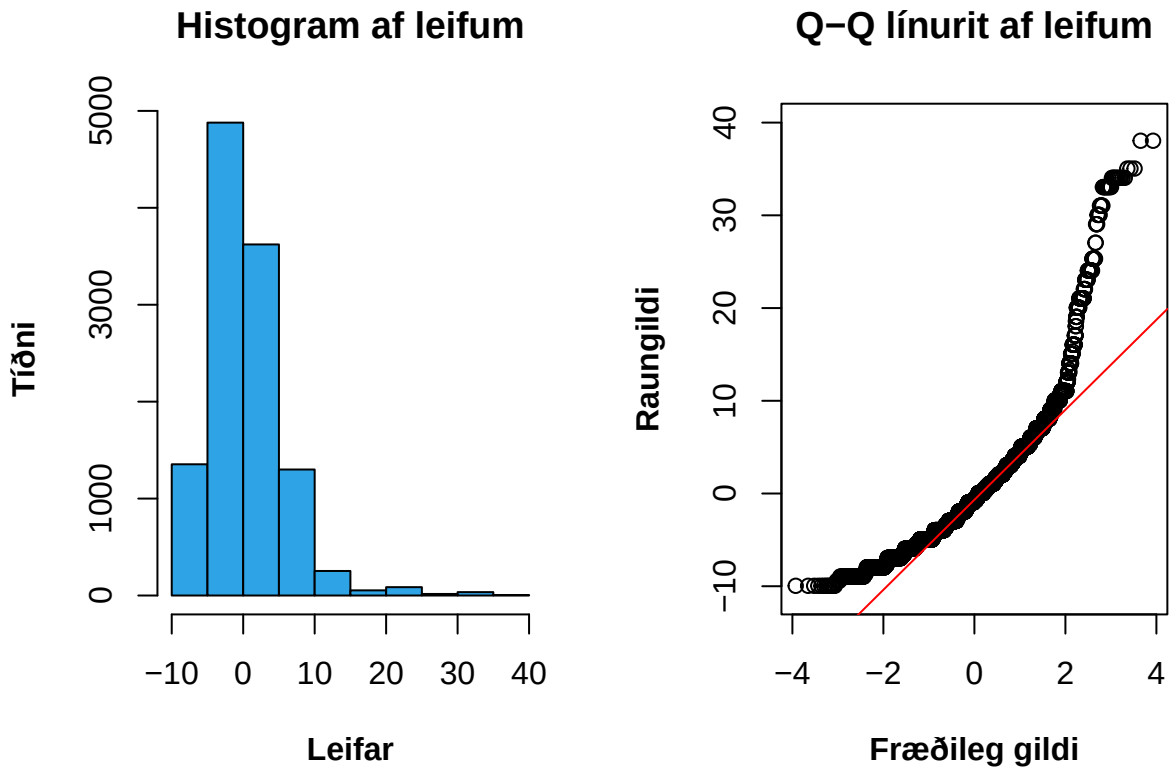
Við höfnum því núlltilgátunni og ályktum að eldsneytistegund hafi áhrif á sparneytni bíla í borgarakstri.

Skoðum dreifinguna

Eftir að hafa framkvæmt ANOVA-próf er mikilvægt að kanna hvort forsendur þess haldist. Sérstaklega þarf að athuga hvort leifarnar séu normaldreifðar og hvort dreifing milli hópa sé svipuð.

```r
leifar <- residuals(anova_prof)

par(mfrow = c(1, 2))

hist(leifar,
     main = "Histogram af leifum",
     xlab = "Leifar",
     ylab = "Tíðni",
     col = "#2ea4e7",
     font.lab = 2)

qqnorm(leifar,
       main = "Q-Q línurit af leifum",
       xlab = "Fræðileg gildi",
       ylab = "Raungildi",
```

```
       ylim = c(-11, 40),
       font.lab = 2
       )
qqline(leifar, col = "red")
```




```
par(mfrow = c(1, 1))
```

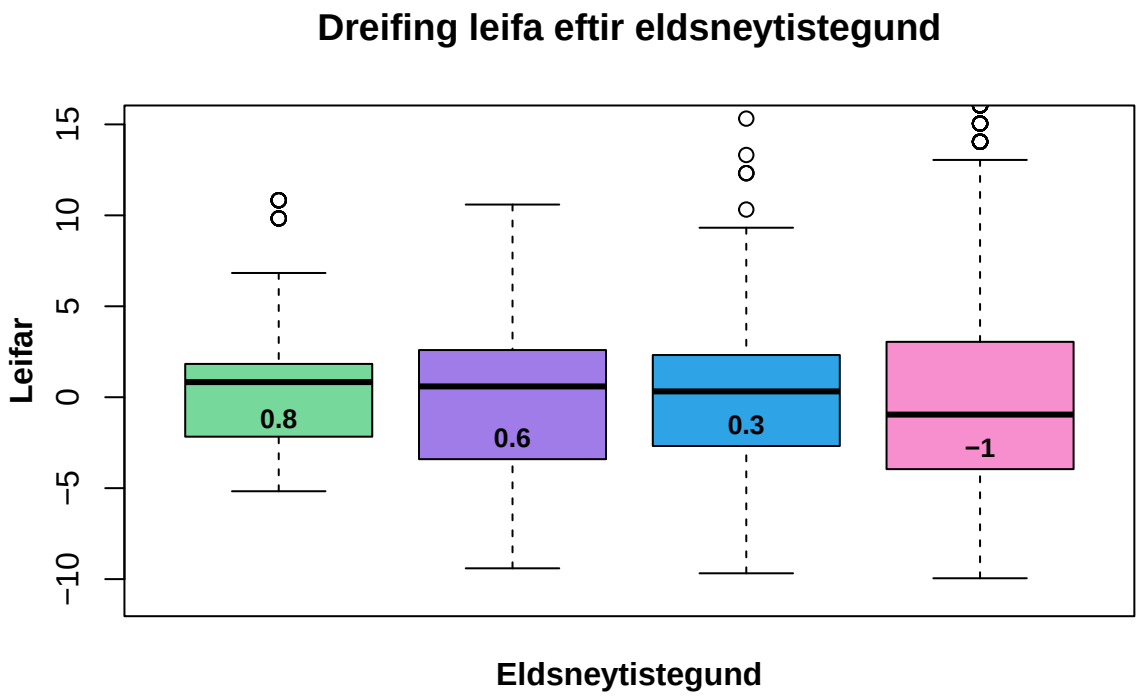
Histogramið sýnir að flestar leifarnar dreifast nálægt núlli en dreifingin er skekkt til hægri.

Í Q-Q plottinu sjáum við að punktarnir fylgja línunni ágætlega í miðjunni en víkja frá í endunum, sérstaklega efst. Þetta bendir til þess að leifarnar séu ekki alveg normaldreifðar, þó meginhluti þeirra fylgi línunni sæmilega.

Það eru ákveðin frávik frá normaldreifingu (sérstaklega í toppnum) og því er rétt að skoða einnig umraðanapróf, t.d. Kruskal–Wallis próf sem gerir ekki ráð fyrir normaldreifingu leifa.

```
boxplot_gildi <- boxplot(leifar ~ data_lidur2$Engine.Fuel.Type,
        xaxt = "n",
        main = "Dreifing leifa eftir eldsneytistegund",
        xlab = "",
        ylab = "",
        col = litir,
        font.lab = 2,
        ylim= c(-11,15))
```

```r
mtext("Eldsneytistegund", side = 1, line = 1, font = 2)
mtext("Leifar", side = 2, line = 2.2, font = 2)

midgildi <- round(boxplot_gildi$stats[3,],1)
hlidrun <- c(-3.6, -4.4, -3.4, -3.4)

text(x = 1:4,
     y = boxplot_gildi$stats[3, ] + hlidrun,
     labels = midgildi,
     pos = 3,
     cex = 0.8,
     font = 2
     )
```



Boxplot-ritið sýnir hvernig leifarnar dreifast eftir eldsneytistegundum. Leifarnar eru munurinn á raunverulegri eldsneytiseyðslu og þeirri sem spáð er samkvæmt ANOVA-líkaninu. Ef forsendur ANOVA-prófsins haldast, ættu leifarnar að vera nálægt núlli, normaldreifðar og með svipaða dreifingu milli hópa. Til að bæta upplausn og læsileika boxplotsins voru útlæg gildi yfir ákveðnu marki falin í myndritinu (en ekki fjarlægð úr gagnasafninu). Auk þess voru töluleg miðgildi sýnd til að auðvelda túlkun á miðlægri dreifingu innan hvers hóps.

Á grafinu má sjá að leifarnar eru mismunandi dreifðar eftir eldsneytistegundum sem bendir til ósamræmis í dreifingu milli hópa. Sérstaklega fyrir “regular unleaded” má sjá fjölda útlægra gilda sem sýna veruleg frávik frá spáðri meðaleyðslu. Einnig er miðgildið ekki alltaf staðsett í miðjunni sem bendir til skekkju í dreifingu. Að lokum má sjá að dreifing leifanna er ekki jöfn milli hópa.

Kruskal-Wallis próf

Þar sem forsendur ANOVA-prófsins haldast ekki fullkomlega framkvæmum við umraðanapróf. Kruskal–Wallis próf er hentugt í þessu samhengi þar sem það gerir ekki ráð fyrir normaldreifingu gagna og ber saman miðgildi hópa í stað meðaltala.

```
library(knitr)
library(kableExtra)

kruskal <- kruskal.test(city.mpg ~ Engine.Fuel.Type, data = data_lidur2)

kruskal_tafla <- data.frame(
  "Próf" = "Kruskal-Wallis",
  "Chi²" = round(kruskal$statistic, 2),
  "Frígráður" = kruskal$parameter,
  "P-gildi" = format.pval(kruskal$p.value, digits = 3, eps = .001)
)

kable(kruskal_tafla, format = "latex", caption = "Niðurstaða Kruskal-Wallis prófs", booktabs = TRUE) %>%
  kable_styling(latex_options = c("striped", "hold_position")) %>%
  column_spec(1:4, width = "3cm")
```

Table 4: Niðurstaða Kruskal–Wallis prófs

	Próf	Chi.	Frígráður	P.gildi
Kruskal-Wallis chi-squared	Kruskal–Wallis	1089.44	3	<0.001

Niðurstöður Kruskal–Wallis prófsins sýna að p-gildið er < 0.001 sem gefur til kynna að miðgildi borgarakstursseyðslu séu marktækt ólík milli a.m.k. einnar eldsneytistegundar og hinna.

Þetta styður niðurstöður ANOVA-prófsins en með minni kröfur um forsendur. Því getum við ályktað með góðri vissu að eldsneytistegund hefur marktæk áhrif á eyðslu bíla í borg.

ANOVA VS Kruskal–Wallis

Bæði ANOVA og Kruskal–Wallis prófin leiddu af sér mjög marktækar niðurstöður (p-gildi < 0.001) sem bendir til þess að það sé munur á meðaleyðslu bíla í borg eftir eldsneytistegund.

ANOVA gerir ráð fyrir að leifarnar séu normaldreifðar og að dreifing sé jöfn milli hópa. Eins og við sáum í greiningu á leifum haldast þessar forsendur ekki fullkomlega í gögnunum okkar, það mátti greina skekkju og útlæg gildi.

Kruskal–Wallis prófið gerir ekki ráð fyrir normaldreifingu og byggir á röðun gagna frekar en hráum gildum. Það er því öruggara þegar forsendur ANOVA eru brotnar.

Það að bæði prófin skili svipuðum niðurstöðum bendir til þess að marktækur munur sé raunverulega til staðar og að ANOVA hafi samt sem áður náð að greina hann þrátt fyrir brotnar forsendur. Þetta styrkir niðurstöðuna og bendir til þess að eldsneytistegund hafi raunveruleg áhrif á meðaleyðslu í borgarakstri.

Bjagar sem geta skekkt niðurstöðurnar

Eldsneytistegundir valdar:

Við völdum aðeins fjórar algengustu eldsneytistegundirnar til að bæta læsileika grafsins. Þetta einfaldaði greininguna en útilokar sjaldgæfari valkosti og getur rýrt dýpt niðurstöðunnar.

Úrtaksbjagi:

Ekki liggur fyrir hvernig gagnasafnið var tekið saman. Ef það byggir eingöngu á tilteknum tegundum bíla (t.d. eingöngu amerískir bílar) eða tilteknum framleiðendum gæti úrtakið verið ófulltrúandi fyrir alla bíla. Slíkur úrtaksbjagi getur skekkt niðurstöðurnar.

Breytur:

Í þessum lið skoðuðum við aðeins samband eldsneytistegundar og eldsneytiseyðslu en höfðum ekki stjórn á öðrum þáttum sem geta haft áhrif svo sem þyngd bíls, vélartækni eða árgerð. Slíkar breytur gætu valdið skekkju í niðurstöðunum og haft áhrif á túlkun.

Skráningarvilla:

Ef einhverjir mæligallar eða rangt skráð gögn leynast í gagnasafninu (t.d. óvenjulega há eða lág MPG gildi) getur það haft áhrif á meðaleyðslu og leitt til skekkju í greiningunni.

Samantekt

Í þessum lið skoðuðum við hvort meðaleyðsla bíla í borgarakstri væri breytileg eftir eldsneytistegund. Við nýttum bæði stöplarit og tölfræðipróf til að meta hvort marktækur munur væri til staðar. Greiningin sýndi að bílar sem nota mismunandi eldsneyti eru með ólíka meðaleyðslu og að munurinn er marktækur samkvæmt bæði ANOVA og Kruskal–Wallis prófum.

Skoðun á leifum sýndi að forsendur ANOVA prófsins haldast ekki að fullu, sérstaklega vegna skekkju og ójafnrar dreifingar milli hópa. Því veitir Kruskal–Wallis prófið traustari niðurstöðu þar sem það byggir ekki á normaldreifingu gagna. Báðar aðferðir bentu þó til sömu niðurstöðu að eldsneytistegund hefur áhrif á sparneytni bíla í borgarakstri.

Einnig var farið yfir mögulega bjaga í gögnunum. Þrátt fyrir það gefa niðurstöðurnar innsýn í hvernig val á eldsneyti getur tengst sparneytni og hönnun bíla. Í raunheimi getur þetta þýtt að kaupendur sem vilja hámarka sparneytni ættu að huga að því hvaða eldsneytistegund bíllinn notar. Til dæmis gæti neytandi sem keyrir mikið í borg valið bíl sem notar „regular unleaded“ frekar en „premium“ ef sparnaður er forgangsatriði.

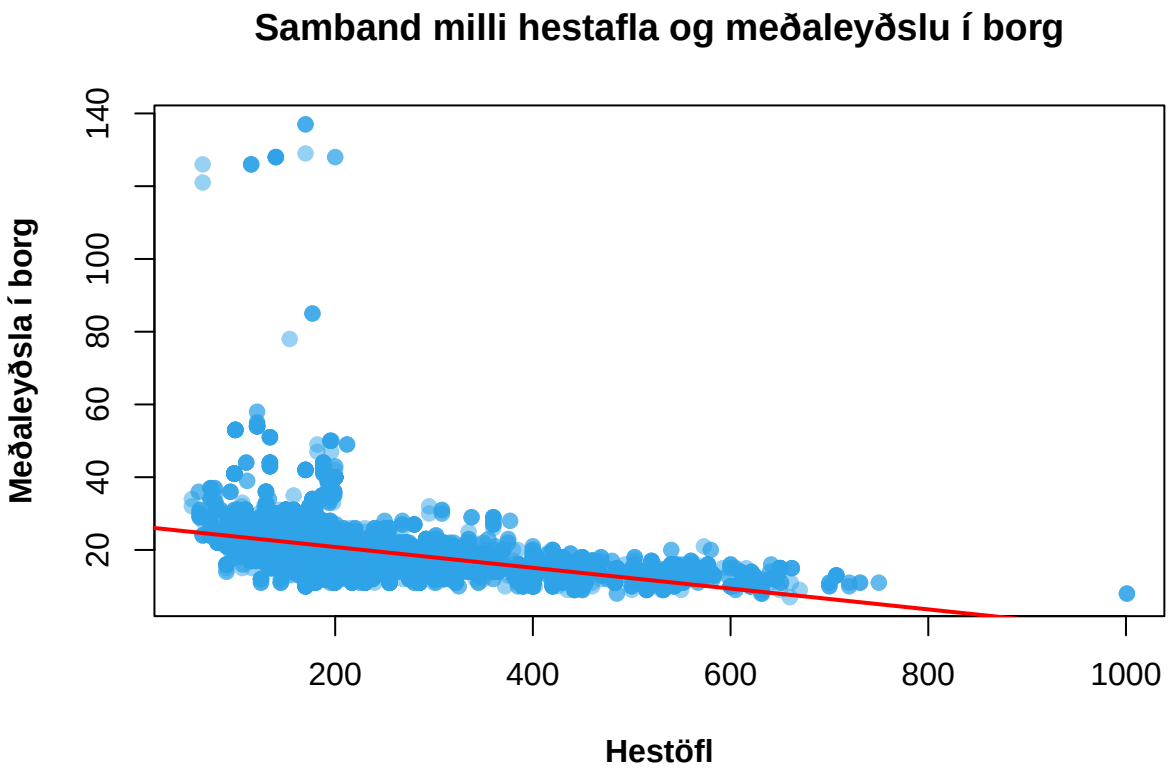
Liður 3 - Tvær talnabreytur

Er fylgni milli hestafla og meðaleyðslu í borg?

Við viljum kanna hvort samband sé á milli hestafla ökutækis og meðaleyðslu í borg. Til þess skoðum við tengsl tveggja samfelldra talnabreytna: *Engine HP* (hestöfl) og *city MPG* (meðaleyðsla í borg).

Við byrjum á því að teikna dreifirit sem sýnir hvernig meðaleyðsla í borg breytist með auknum hestöflum. Þetta gefur okkur sjónræna vísbendingu um hvort línulegt samband sé á milli breytanna og hvort hægt sé að draga ályktanir um tengsl afls og sparneytni í daglegum borgarakstri.

```
data_lidur3 <- subset(data, !is.na(Engine.HP) & !is.na(city.mpg))

plot(data_lidur3$Engine.HP, data_lidur3$city.mpg,
     main = "Samband milli hestafla og meðaleyðslu í borg",
     xlab = "Hestöfl",
     ylab = "Meðaleyðsla í borg",
     pch = 19,
     col = rgb(0.18, 0.64, 0.91, 0.5),
     font.lab = 2
     )

abline(lm(city.mpg ~ Engine.HP, data = data_lidur3), col = "red", lwd = 2)
```



Dreifiritið hér að ofan sýnir sambandið milli hestafla ökutækis og eldsneytiseyðslu í borgarakstri. Hver punktur táknar einn bíl þar sem x-ásinn sýnir hestöfl og y-ásinn sýnir eldsneytiseyðslu í MPG.

Á myndinni má greinilega sjá neikvætt samband á milli breytanna, þegar hestöflin aukast, minnkar eldsneytiseyðslan almennt, þ.e. bílarnir verða minna sparneytnir. Rauða línan táknar línulega aðhvarfslínu sem gefur til kynna almenna stefnu sambandsins og sýnir að bílar með öflugri vélar eyða að jafnaði meira eldsneyti í borgarakstri.

Þó að sambandið sé greinilegt í heildina, er töluverður breytileiki milli einstakra bíla með sambærileg hestöfl. Einnig má sjá að sumir bílar með lítið afl skera sig úr með óvenju mikla eldsneytiseyðslu sem gæti bent til utanaðkomandi áhrifa eins og stærðar, þyngdar eða afkastagetu ökutækis.

Fylgnistuðull

Til að meta sambandið á milli hestafla og eldsneytiseyðslu í borg könnum við fylgnistuðul þeirra. Ef fylgnistuðullinn er neikvæður, bendir það til þess að því fleiri hestöfl sem bíll hefur, því minni er sparneytni hans í borgarakstri (þ.e. hann eyðir meira eldsneyti). Sé stuðullinn nálægt núlli, bendir það til þess að lítið eða ekkert línulegt samband sé milli breytanna.

Við reiknum Pearson-fylgnistuðul milli breytanna *Engine HP* og *city.mpg*:

```
data_lidur3 <- subset(data, !is.na(Engine.HP) & !is.na(city.mpg))

fylgni <- cor(data_lidur3$Engine.HP, data_lidur3$city.mpg)

fylgni
```

```
## [1] -0.4393713
```

Fylgnistuðullinn sem við fáum er neikvæður (-0.44) sem styður við niðurstöðuna sem sást í dreifiritinu. Því fleiri hestöfl sem bíll hefur, því minni verður eldsneytiseyðsla hans mæld í MPG (sem í raun þýðir aukin eyðsla). Þetta styrkir grun okkar um að afl hafi áhrif á sparneytni bílsins í borg.

Tilgátupróf

Til að kanna hvort marktækt línulegt samband sé á milli hestafla og eldsneytiseyðslu í borg, framkvæmum við hefðbundið tilgátupróf fyrir Pearson-fylgni. Þetta próf metur hvort fylgnistuðullinn sé marktækt frábrugðinn núlli (engin fylgni).

Við notum cor.test() fallið í R til að reikna bæði fylgni og framkvæma tilgátuprófið:

```
cor_test <- cor.test(data_lidur3$Engine.HP, data_lidur3$city.mpg)

library(knitr)
library(kableExtra)

cor_tafla <- data.frame(
  "Próf" = "Pearson fylgnipróf",
  "Fylgni r" = round(cor_test$estimate, 3),
  "Frígráður df" = round(cor_test$parameter, 0),
  "P-gildi" = format.pval(cor_test$p.value, digits = 3, eps = 0.001)
  )

kable(cor_tafla, format = "latex", caption = "Niðurstaða Pearson fylgniprófs", booktabs = TRUE) %>%
  kable_styling(latex_options = c("striped", "hold_position")) %>%
  column_spec(1:4, width = "3cm")
```

Table 5: Niðurstaða Pearson fylgniprófs

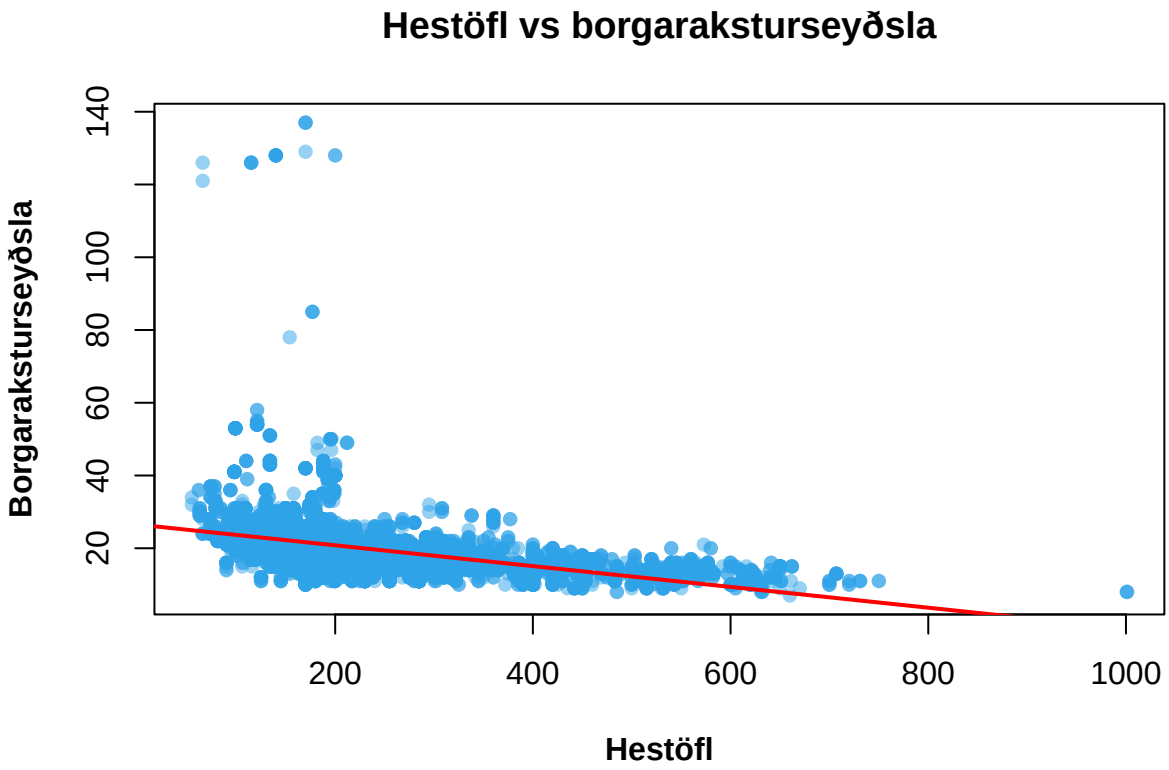
	Próf	Fylgni.r	Frígráður.df	P.gildi
cor	Pearson fylgnipróf	-0.439	11843	<0.001

Niðurstöðurnar sýna að Pearson fylgnistuðullinn er neikvæður sem bendir til þess að eftir því sem hestöfl bílsins aukast, þá lækkar meðaleyðsla í borg. Þetta þýðir að öflugri bílar eru almennt minna sparneytnir í borgarakstri. Þar sem p-gildið er < 0.001 getum við með mikilli vissu hafnað núlltilgátunni um að ekkert samband sé á milli hestafla og eldsneytiseyðslu. Við ályktum því að marktækt neikvætt línulegt samband sé á milli þessara tveggja breyta. Þessar niðurstöður styðja við þá hugmynd að öflugari vélar sem framleiða meiri hestafla, séu yfirleitt tengdar hærri eldsneytiseyðslu, sérstaklega við akstur í borg þar sem hröðun og þyngd hafa meiri áhrif.

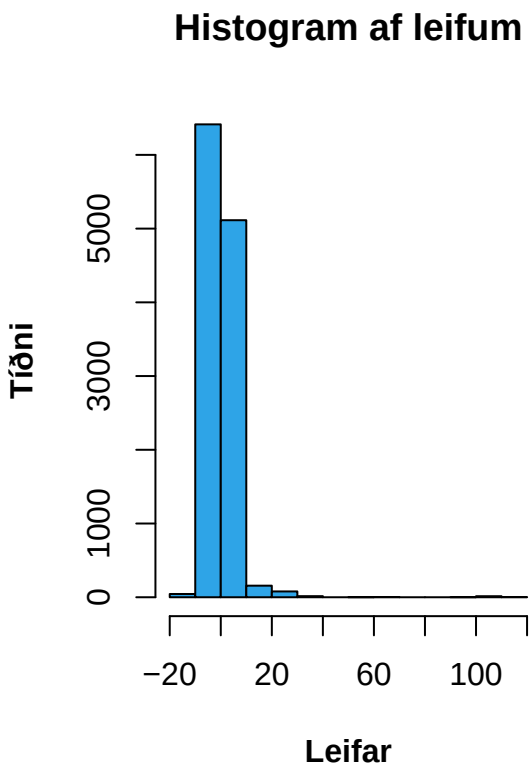
Til að kanna hvort forsendur Pearson-fylgniprófsins haldist teiknum við dreifirit með línu sem sýnir línulegt samband og síðan skoðum við normaldreifingu leifa með histogram og Q-Q plotti.

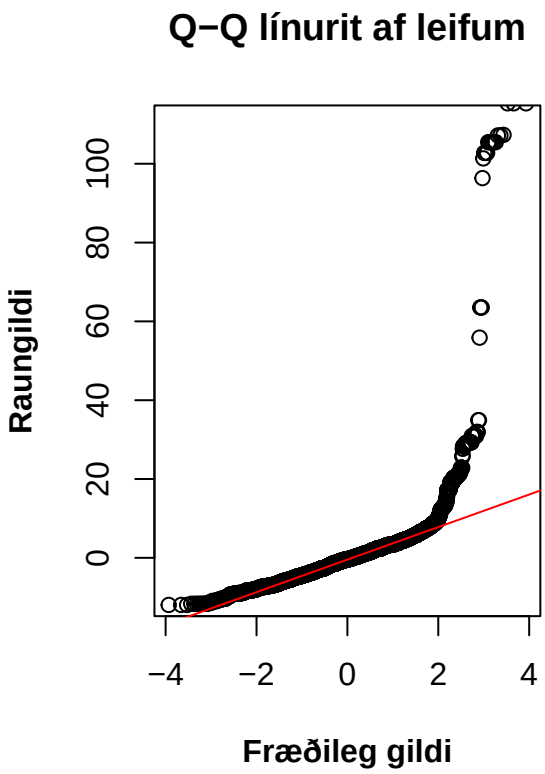
```
lína <- lm(city.mpg ~ Engine.HP, data = data_lidur3)
leifar <- residuals(lína)

plot(data_lidur3$Engine.HP, data_lidur3$city.mpg,
     main = "Hestöfl vs borgaraksturseyðsla",
     xlab = "Hestöfl",
```

```
     ylab = "Borgaraksturseyðsla",
     col = rgb(0.18, 0.64, 0.91, 0.5),
     pch = 16,
     font.lab = 2
    )

abline(lína, col = "red", lwd = 2)
```




```
par(mfrow = c(1, 2))

hist(leifar,
     main = "Histogram af leifum",
     xlab = "Leifar",
     ylab = "Tíðni",
     col = "#2ea4e7",
     font.lab = 2)

qqnorm(leifar,
       main = "Q-Q línurit af leifum",
       xlab = "Fræðileg gildi",
       ylab = "Raungildi",
       ylim = c(-10,110),
       font.lab = 2)
qqline(leifar, col = "red")
```






```
par(mfrow = c(1, 1))
```

Út frá dreifiritinu sést að sambandið á milli hestafla og borgaraksturseyðslu virðist vera línulegt. Leifarnar dreifast nokkuð eðlilega í histograminu og punktarnir fylgja Q-Q línunni nokkuð vel, þó örlítil frávik sjáist á endunum. Það bendir til þess að normaldreifing leifa haldist nokkuð vel, sérstaklega þar sem úrtakið er stórt. Því teljum við að forsendur Pearson-fylgniprófsins séu í aðalatriðum uppfylltar.

Bjagar sem geta skekkt niðurstöðurnar

Eldsneytistegundir valdar:

Við völdum aðeins fjórar algengustu eldsneytistegundirnar til að bæta læsileika grafsins. Þetta einfaldaði greininguna en útilokar sjaldgæfari valkosti og getur rýrt dýpt niðurstöðunnar.

Úrtaksbjagi:

Ekki liggur fyrir hvernig gagnasafnið var tekið saman. Ef það byggir eingöngu á tilteknum tegundum bíla (t.d. eingöngu amerískir bílar) eða tilteknum framleiðendum gæti úrtakið verið ófulltrúandi fyrir alla bíla. Slíkur úrtaksbjagi getur skekkt niðurstöðurnar.

Skráning gagna:

Gögn um eldsneytiseyðslu og hestafla byggja á skráðum upplýsingum frá framleiðendum eða prófunum sem geta verið framkvæmdar við mismunandi aðstæður. Ef aðferðir við mælingar eru ekki stöðlaðar milli bílaframleiðenda getur það leitt til ónákvæmni og skekkt niðurstöður greiningarinnar.

Samantekt

Í þessum hluta skoðuðum við sambandið milli hestafla bíla og eldsneytiseyðslu í borgarakstri. Dreifiritið sýndi greinilegt neikvætt línulegt samband, því meiri sem hestöflin eru, því minni er sparneytni bílsins. Pearson fylgnistuðullinn var -0.44 og p-gildið < 0.001 sem gefur til kynna marktækt neikvætt samband milli breytanna.

Forsendur tilgátuprófsins voru í aðalatriðum uppfylltar þar sem dreifing leifa var nokkuð eðlileg og sambandið virtist línulegt. Því getum við ályktað að öflugri bílar eyði marktækt meira eldsneyti í borg en sparneytnari bílar með minna afl. Þó ber að hafa í huga mögulega bjaga eins og úrtaksbjaga og óstöðluð mælingaskilyrði sem geta haft áhrif á alhæfingargildi niðurstaðna.

Þetta endurspeglar mikilvægt val neytenda á milli afkasta og sparnaðar. Þeir sem leggja áherslu á minni eldsneytiskostnað eða minni losun velja líklega bíla með minni vélum, sérstaklega í borgaraðstæðum þar sem hraði og afköst skipta minna máli. Einnig geta niðurstöðurnar haft áhrif á hönnun og markaðssetningu bíla fyrir mismunandi markhópa svo sem fjölskyldubíla gegn sportbílum.

Lokaorð

Í þessu verkefni var leitast við að svara þremur spurningum um samband bílaeiginleika og eldsneytisnotkun. Fyrsti hluti sýndi að það er samband milli gerðar bíls og þess hvaða eldsneyti hann notar. Í öðrum hluta kom fram að eyðsla í borg er mismunandi eftir eldsneytistegundum, og í þriðja hluta sást að því fleiri hestöfl sem bíll hefur, því minni er sparneytnin í borg.

Niðurstöðurnar gefa innsýn í hvernig ákveðnir tæknilegir þættir eins og afl og eldsneyti, tengjast daglegri notkun bíla. Þær gætu nýst bæði neytendum sem vilja velja sparneytnari bíla og framleiðendum sem eru að hanna bíla fyrir mismunandi markhópa. Þó þarf að hafa í huga að gagnasafnið nær ekki yfir alla bíla og sum gildi voru einfölduð.