**Workshop4**

survery.program1 <-read.csv("survey\_messy.csv", header = TRUE)

survery.program1 <-read.csv("survey\_messy.csv", header = TRUE)

View(survery.program1)

str(survery.program1)

tv.hours.messy <- survery.program1$TVhours

as.character(tv.hours.messy)[1:30]

as.numeric(as.character(tv.hours.messy))[1:30]

typeof(as.numeric(as.character(tv.hours.messy)))

tv.hours.strings <- as.character(tv.hours.messy)

tv.hours.strings

gsub("[^0-9]","",tv.hours.strings)

tv.hours.messy

tv.hours.clean <- as.numeric(gsub("[^0-9]","",tv.hours.strings))

tv.hours.clean

survery.program1 <- read.csv("survey\_messy.csv", header=TRUE, stringsAsFactors = FALSE)

str(survery.program1)

survey <- transform(survery.program1, TVhours = as.numeric(gsub("[^0-9.]", "", TVhours)))

str(survey)

**Workshop5.**

This problem uses the Cars93 dataset from the MASS package.

(1)Use `qplot` to create a scatterplot with Price on the y-axis and EngineSize on the `x-axis`.

qplot(x = EngineSize, y = Price, data = Cars93)

(2) Repeat part (a) using the `ggplot` function and `geom\_point()` layer.

ggplot(Cars93, aes(x = EngineSize, y = Price)) + geom\_point()

(1)Repeat part (b), but this time specifying that the `color` mapping should depend on `Type` and the `shape` mapping should depend on `DriveTrain`.

ggplot(Cars93, aes(x = EngineSize, y = Price, colour = Type, shape = DriveTrain)) + geom\_point()

**Workshop6.**

## Including Plots

You can also embed plots, for example:

plot(pressure)

#### Plotting the Cars93 data

This problem uses the Cars93 dataset from the MASS package.

**(a)** Use qplot to create a scatterplot with Price on the y-axis and EngineSize on the x-axis.

qplot(x = EngineSize, y = Price, data = Cars93)

**(b)** Repeat part (a) using the ggplot function and geom\_point() layer.

ggplot(Cars93, aes(x = EngineSize, y = Price)) + geom\_point()

**(c)** Repeat part (b), but this time specifying that the color mapping should depend on Type and the shape mapping should depend on DriveTrain.

ggplot(Cars93, aes(x = EngineSize, y = Price, colour = Type, shape = DriveTrain)) + geom\_point()

# Create data for the graph.

x <- c(21, 62, 10,53) labels <- c("London","New York","Singapore","Mumbai")

piepercent<- round(100\*x/sum(x), 1)

# Give the chart file a name.

png(file = "city\_percentage\_legends.jpg")

# Plot the chart.

pie(x, labels = piepercent, main = "City pie chart",col = rainbow(length(x))) legend("topright", c("London","New York","Singapore","Mumbai"), cex = 0.8, fill = rainbow(length(x)))

# Save the file.

dev.off()

**Workshop7**

**Is the data normal?**

**(a)** Construct histograms of MPG.highway, one plot for each Origin category.

qplot(x = MPG.highway, data = Cars93, facets = ~Origin, geom = "histogram", fill = Origin, binwidth = 2)

**(b)** Does the data look to be normally distributed?

The histograms don't really look normally distributed, so we might be better off using the non-parametric test.

**(c)** Construct qqplots of MPG.highway, one plot for each Origin category. Overlay a line on each plot using with qqline()function.

par(mfrow = c(1,2))

# USA cars

with(Cars93, qqnorm(MPG.highway[Origin == "USA"]))

with(Cars93, qqline(MPG.highway, col = "blue"))

# Foreign cars

with(Cars93, qqnorm(MPG.highway[Origin == "non-USA"]))

with(Cars93, qqline(MPG.highway, col = "blue"))

**(d)** Does the data look to be normally distributed?

The non-USA MPG.highway data looks very far from normally distributed.

**Testing means between two groups**

**(a)** Using the Cars93 data and the t.test() function, run a t-test to see if average MPG.highway is different between US and non-US vehicles.

Try doing this both using the formula style input and the x, y style input.

# Formula version

mpg.t.test <- t.test(MPG.highway ~ Origin, data = Cars93)

mpg.t.test

**at confident level = 90%**

mpg.t.test <- t.test(MPG.highway ~ Origin, data = Cars93, conf.level = 0.90)

mpg.t.test

**x, y version**

with(Cars93, t.test(x = MPG.highway[Origin == "USA"], y = MPG.highway[Origin == "non-USA"]))

\*\*(b)\*\* What is the confidence interval for the difference?

```{r}

mpg.t.test$conf.int

**(c)** Repeat part (a) using the wilcox.test() function.

mpg.wilcox.test <- wilcox.test(MPG.highway ~ Origin, data = Cars93)

mpg.wilcox.test