ACT05 - Data Mining

Team 7 | Captain: Nadine Rose | Members: Kyle Scott, Lakshya Rathore, Bailey LaRea

March 1, 2022

Problem 1 - Programming

One - Read Data

```r
# read in the csv
library(readr)
hd <- read_csv("HeartDisease.csv")
```

Two - Univariate Statistics

```r
library(psych)
library(kableExtra)
library(dplyr)

# select only the needed columns and gather statistics
df <- select(hd, -c(names, famhist, chd))
tbl <- describe(df, skew = TRUE)
# calculate vairance
variance <- (tbl$sd)^2
variance <- round(variance, digits = 3)
# calculate first quartile
first_quartile <- summarize_all(df, ~ quantile(.x, 0.25))
first_quartile <- as.numeric(first_quartile[1,])
# calculate third quartile
third_quartile <- summarize_all(df, ~quantile(.x, 0.75))
third_quartile <- as.numeric(third_quartile[1,])
# bind extra statistics to existing tbl
tbl <- cbind(tbl, variance, first_quartile, third_quartile)
# reorder tbl
tbl <- tbl[, -c(1, 6, 7)]
tbl <- tbl[, c(2, 4, 8, 9, 3, 11, 10, 12, 13)]
# print tbl
kable(tbl, caption = "Univariate Statistics for Heart Disease Data", linesep = "\\addlinespace", digits = 
      booktabs = T, format = 'pandoc')
```

Table 1: Univariate Statistics for Heart Disease Data

	mean	median	skew	kurtosis	sd	variance	se	first_quartile	third_quartile
sbp	138.327	134.000	1.173	1.729	20.496	420.099	0.954	124.000	148.000
tobacco	3.636	2.000	2.066	5.852	4.593	21.096	0.214	0.053	5.500
ldl	4.740	4.340	1.305	2.807	2.071	4.289	0.096	3.282	5.790
adiposity	25.407	26.115	-0.213	-0.714	7.781	60.539	0.362	19.775	31.227
typea	53.104	53.000	-0.344	0.437	9.818	96.384	0.457	47.000	60.000
obesity	26.044	25.805	0.899	2.196	4.214	17.755	0.196	22.985	28.497
alcohol	17.044	7.510	2.298	6.298	24.481	599.322	1.139	0.510	23.892
age	42.816	45.000	-0.379	-1.027	14.609	213.422	0.680	31.000	55.000

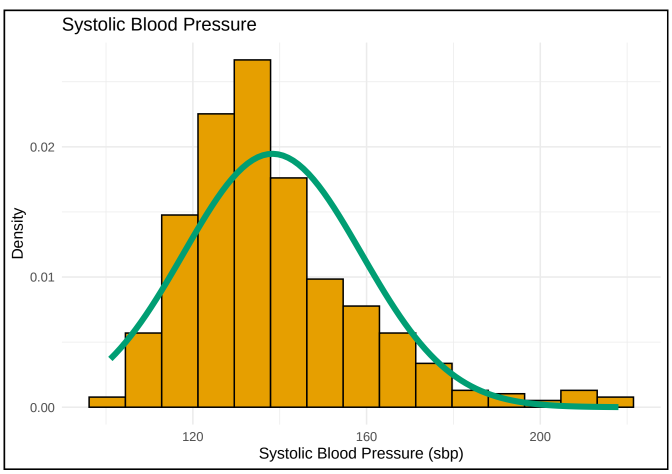
Three - Histograms

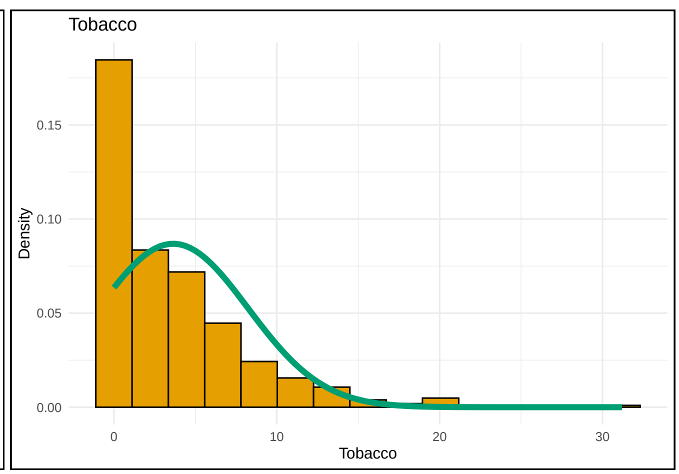
```r
library(ggplot2)
library(ggthemes)

plt_sbp <- ggplot(data = hd, aes(x = sbp)) +
  geom_histogram(aes(y = ..density..), col = "black", fill = "#E69F00", bins = 15) +
  stat_function(fun = dnorm, args = list(mean = mean(hd$sbp),
                                         sd = sd(hd$sbp)),
                col = "#009E73", size = 2) +
  ggtitle(label = "Systolic Blood Pressure") +
  xlab("Systolic Blood Pressure (sbp)") +
  ylab("Density") +
  theme_minimal() +
  theme(plot.background = element_rect(colour = "black", fill=NA, size=1))

plt_sbp

plt_tobacco <- ggplot(data = hd, aes(x = tobacco)) +
  geom_histogram(aes(y = ..density..), col = "black", fill = "#E69F00", bins = 15) +
  stat_function(fun = dnorm, args = list(mean = mean(hd$tobacco),
                                         sd = sd(hd$tobacco)),
```

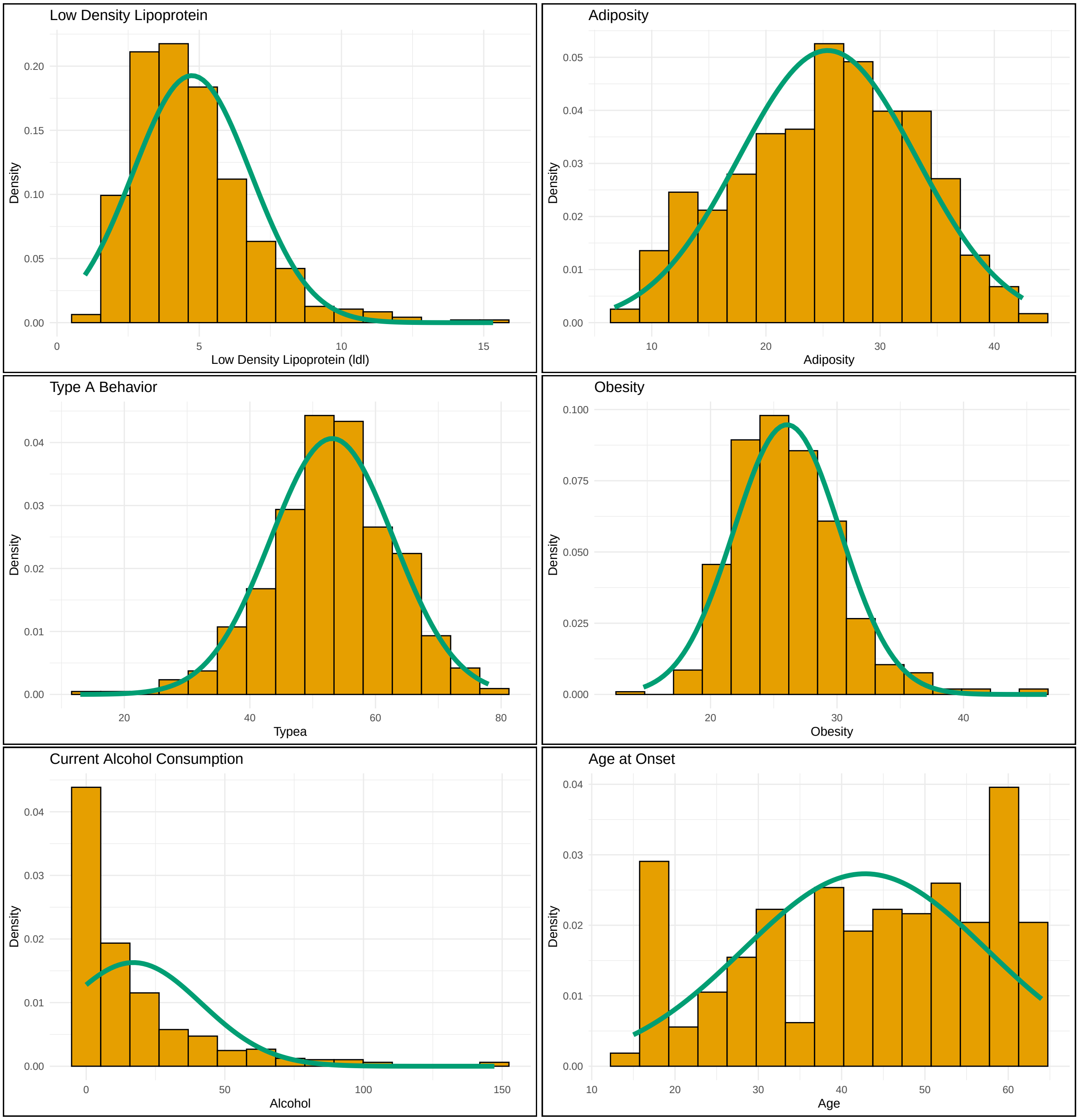
```
                col = "#009E73", size = 2) +
  ggtitle(label = "Tobacco") +
  xlab("Tobacco") +
  ylab("Density") +
  theme_minimal() +
  theme(plot.background = element_rect(colour = "black", fill=NA, size=1))

plt_tobacco

plt_ldl <- ggplot(data = hd, aes(x = ldl)) +
  geom_histogram(aes(y = ..density..), col = "black", fill = "#E69F00", bins = 15) +
  stat_function(fun = dnorm, args = list(mean = mean(hd$ldl),
                                         sd = sd(hd$ldl)),
                col = "#009E73", size = 2) +
  ggtitle(label = "Low Density Lipoprotein") +
  xlab("Low Density Lipoprotein (ldl)") +
  ylab("Density") +
  theme_minimal() +
  theme(plot.background = element_rect(colour = "black", fill=NA, size=1))

plt_ldl

plt_adiposity <- ggplot(data = hd, aes(x = adiposity)) +
  geom_histogram(aes(y = ..density..), col = "black", fill = "#E69F00", bins = 15) +
  stat_function(fun = dnorm, args = list(mean = mean(hd$adiposity),
                                         sd = sd(hd$adiposity)),
                col = "#009E73", size = 2) +
  ggtitle(label = "Adiposity") +
  xlab("Adiposity") +
  ylab("Density") +
  theme_minimal() +
  theme(plot.background = element_rect(colour = "black", fill=NA, size=1))

plt_adiposity
```

```r
plt_typea <- ggplot(data = hd, aes(x = typea)) +
  geom_histogram(aes(y = ..density..), col = "black", fill = "#E69F00", bins = 15) +
  stat_function(fun = dnorm, args = list(mean = mean(hd$typea),
                                         sd = sd(hd$typea)),
                col = "#009E73", size = 2) +
  ggtitle(label = "Type A Behavior") +
  xlab("Typea") +
  ylab("Density") +
  theme_minimal() +
  theme(plot.background = element_rect(colour = "black", fill=NA, size=1))

plt_typea

plt_obesity <- ggplot(data = hd, aes(x = obesity)) +
  geom_histogram(aes(y = ..density..), col = "black", fill = "#E69F00", bins = 15) +
  stat_function(fun = dnorm, args = list(mean = mean(hd$obesity),
                                         sd = sd(hd$obesity)),
                col = "#009E73", size = 2) +
  ggtitle(label = "Obesity") +
  xlab("Obesity") +
  ylab("Density") +
  theme_minimal() +
  theme(plot.background = element_rect(colour = "black", fill=NA, size=1))

plt_obesity

plt_alcohol <- ggplot(data = hd, aes(x = alcohol)) +
  geom_histogram(aes(y = ..density..), col = "black", fill = "#E69F00", bins = 15) +
  stat_function(fun = dnorm, args = list(mean = mean(hd$alcohol),
                                         sd = sd(hd$alcohol)),
                col = "#009E73", size = 2) +
  ggtitle(label = "Current Alcohol Consumption") +
  xlab("Alcohol") +
```

```r
  ylab("Density") +
  theme_minimal() +
  theme(plot.background = element_rect(colour = "black", fill=NA, size=1))

plt_alcohol

plt_age <- ggplot(data = hd, aes(x = age)) +
  geom_histogram(aes(y = ..density..), col = "black", fill = "#E69F00", bins = 15) +
  stat_function(fun = dnorm, args = list(mean = mean(hd$age),
                                          sd = sd(hd$age)),
                col = "#009E73", size = 2) +
  ggtitle(label = "Age at Onset") +
  xlab("Age") +
  ylab("Density") +
  theme_minimal() +
  theme(plot.background = element_rect(colour = "black", fill=NA, size=1))

plt_age
```





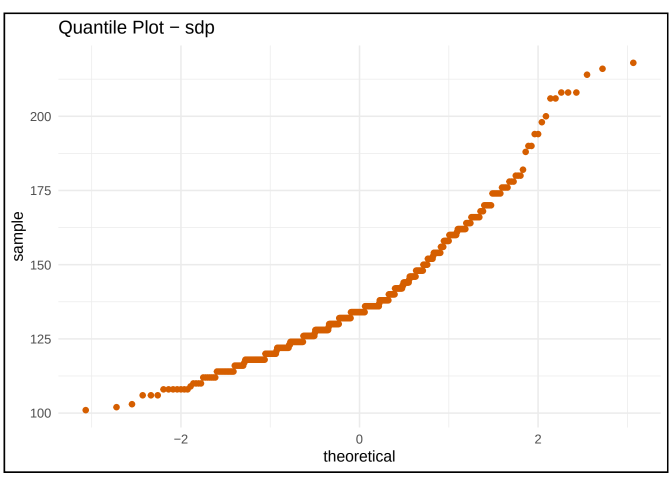


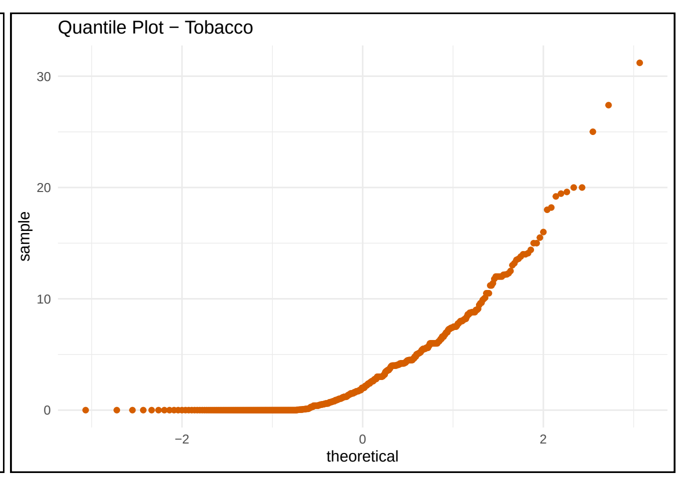
Four - Quantile Plots

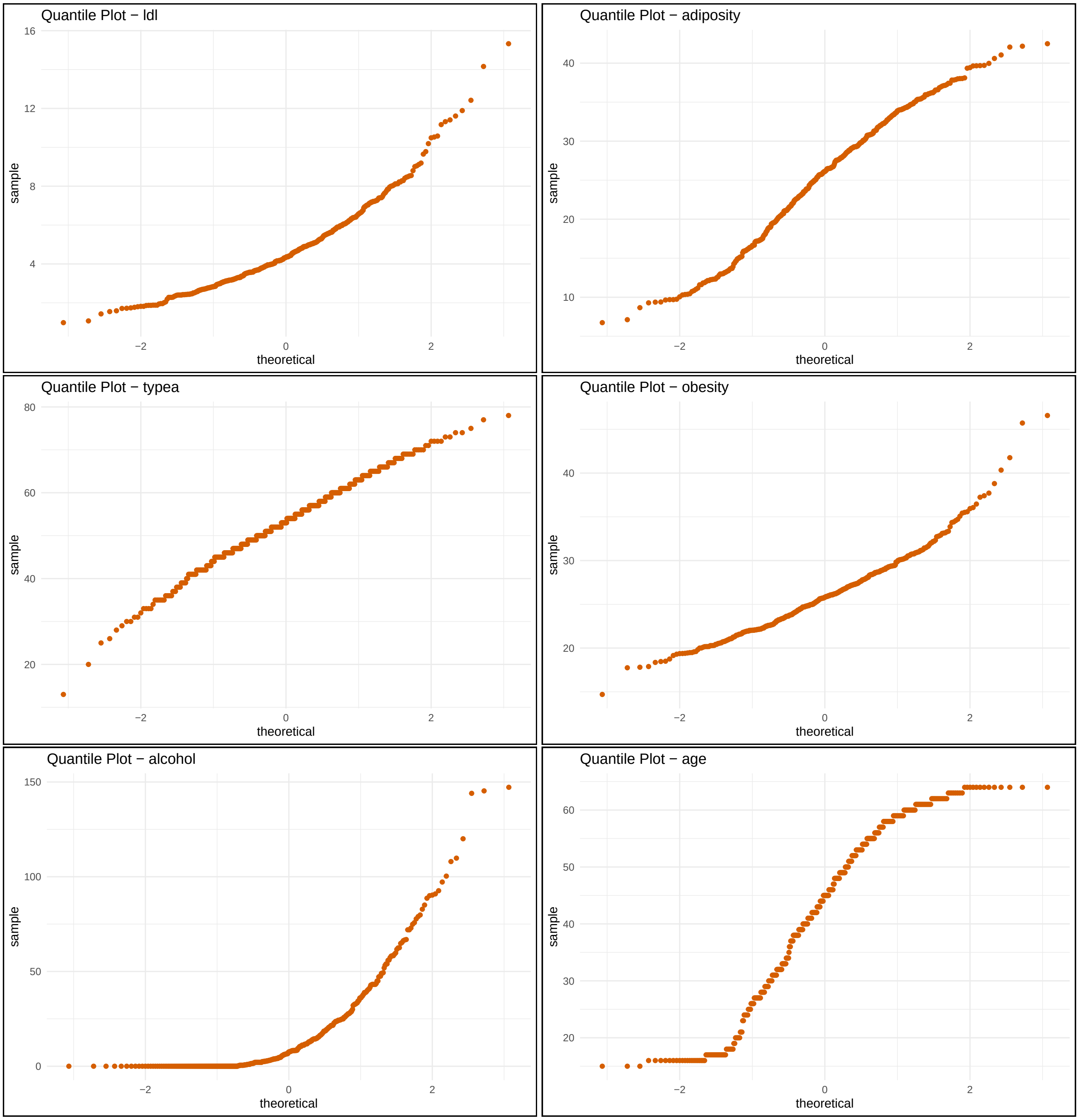
```
qplt_sdp <- ggplot(data = hd, aes(sample = sbp)) +
  stat_qq(col = "#D55E00") +
  ggtitle("Quantile Plot - sdp") +
  theme_minimal() +
  theme(plot.background = element_rect(colour = "black", fill=NA, size=1))
```

```
qplt_sdp

qplt_tobacco <- ggplot(data = hd, aes(sample = tobacco)) +
  stat_qq(col = "#D55E00") +
  ggtitle("Quantile Plot - Tobacco") +
  theme_minimal() +
  theme(plot.background = element_rect(colour = "black", fill=NA, size=1))

qplt_tobacco

qplt_ldl <- ggplot(data = hd, aes(sample = ldl)) +
  stat_qq(col = "#D55E00") +
  ggtitle("Quantile Plot - ldl") +
  theme_minimal() +
  theme(plot.background = element_rect(colour = "black", fill=NA, size=1))

qplt_ldl

qplt_adiposity <- ggplot(data = hd, aes(sample = adiposity)) +
  stat_qq(col = "#D55E00") +
  ggtitle("Quantile Plot - adiposity") +
  theme_minimal() +
  theme(plot.background = element_rect(colour = "black", fill=NA, size=1))

qplt_adiposity

qplt_typea <- ggplot(data = hd, aes(sample = typea)) +
  stat_qq(col = "#D55E00") +
  ggtitle("Quantile Plot - typea") +
  theme_minimal() +
  theme(plot.background = element_rect(colour = "black", fill=NA, size=1))

qplt_typea
```

```r
qplt_obesity <- ggplot(data = hd, aes(sample = obesity)) +
  stat_qq(col = "#D55E00") +
  ggtitle("Quantile Plot - obesity") +
  theme_minimal() +
  theme(plot.background = element_rect(colour = "black", fill=NA, size=1))

qplt_obesity

qplt_alcohol <- ggplot(data = hd, aes(sample = alcohol)) +
  stat_qq(col = "#D55E00") +
  ggtitle("Quantile Plot - alcohol") +
  theme_minimal() +
  theme(plot.background = element_rect(colour = "black", fill=NA, size=1))

qplt_alcohol

qplt_age <- ggplot(data = hd, aes(sample = age)) +
  stat_qq(col = "#D55E00") +
  ggtitle("Quantile Plot - age") +
  theme_minimal() +
  theme(plot.background = element_rect(colour = "black", fill=NA, size=1))

qplt_age
```







Five - Logistic Regression Model

```
logit_model <- glm(formula = chd ~ . - names - famhist, family = binomial(link = "logit"), data = hd)

summary(logit_model)
```

```
##
## Call:
```

```
## glm(formula = chd ~ . - names - famhist, family = binomial(link = "logit"),
##     data = hd)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0519  -0.8392  -0.4681   0.9825   2.4535
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.066864   1.271443  -4.772 1.83e-06 ***
## sbp          0.005641   0.005611   1.005 0.314721
## tobacco      0.072716   0.026326   2.762 0.005742 **
## ldl          0.192492   0.059429   3.239 0.001199 **
## adiposity    0.017066   0.028433   0.600 0.548355
## typea        0.040467   0.012078   3.350 0.000807 ***
## obesity     -0.057931   0.042980  -1.348 0.177703
## alcohol      0.001446   0.004403   0.328 0.742627
## age          0.050650   0.011766   4.305 1.67e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 596.11  on 461  degrees of freedom
## Residual deviance: 488.89  on 453  degrees of freedom
## AIC: 506.89
##
## Number of Fisher Scoring iterations: 4
```

Six - Power Transformations

```
library(forecast)
pwr_sbp <- (hd$sbp) ^ (-2)
pwr_tobacco <- (hd$tobacco) ^ (0.4)
pwr_ldl <- (hd$ldl) ^ (0.1)
```

```r
pwr_obesity <- (hd$obesity) ^ (-0.4)
pwr_alcohol <- (hd$alcohol) ^ (0.4)
```

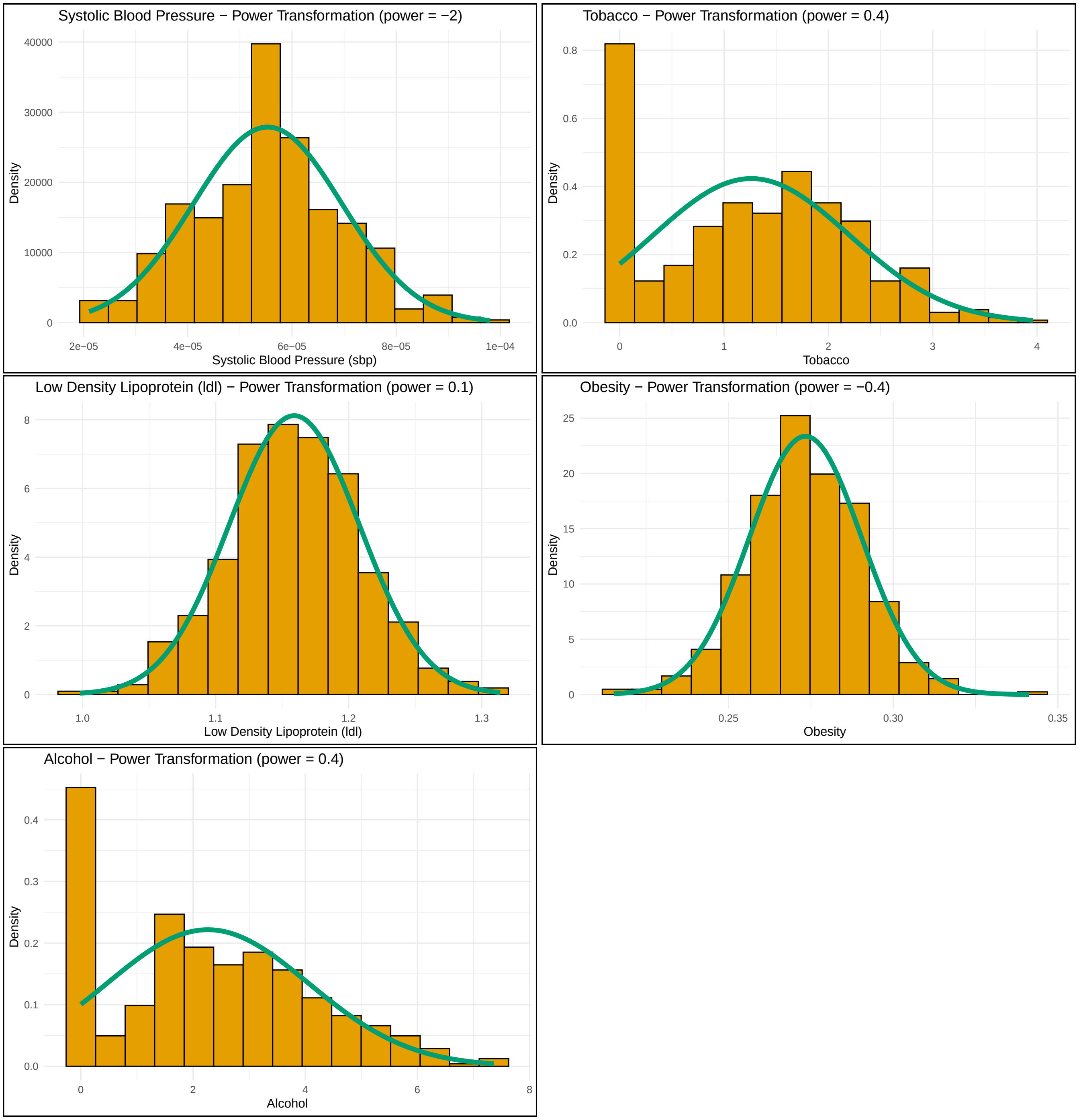
```r
hd <- cbind(hd, pwr_sbp, pwr_tobacco, pwr_ldl, pwr_obesity, pwr_alcohol)
```

Seven - Power Transformation Histograms

```r
library(ggplot2)
library(ggthemes)

plt_pwrsbp <- ggplot(data = hd, aes(x = pwr_sbp)) +
  geom_histogram(aes(y = ..density..), col = "black", fill = "#E69F00", bins = 15) +
  stat_function(fun = dnorm, args = list(mean = mean(hd$pwr_sbp),
                                         sd = sd(hd$pwr_sbp)),
                col = "#009E73", size = 2) +
  ggtitle(label = "Systolic Blood Pressure - Power Transformation (power = -2)") +
  xlab("Systolic Blood Pressure (sbp)") +
  ylab("Density") +
  theme_minimal() +
  theme(plot.background = element_rect(colour = "black", fill=NA, size=1))

plt_pwrsbp

plt_pwrtobacco <- ggplot(data = hd, aes(x = pwr_tobacco)) +
  geom_histogram(aes(y = ..density..), col = "black", fill = "#E69F00", bins = 15) +
  stat_function(fun = dnorm, args = list(mean = mean(hd$pwr_tobacco),
                                         sd = sd(hd$pwr_tobacco)),
                col = "#009E73", size = 2) +
  ggtitle(label = "Tobacco - Power Transformation (power = 0.4)") +
  xlab("Tobacco") +
  ylab("Density") +
  theme_minimal() +
  theme(plot.background = element_rect(colour = "black", fill=NA, size=1))

plt_pwrtobacco
```

```
plt_pwrldl <- ggplot(data = hd, aes(x = pwr_ldl)) +
  geom_histogram(aes(y = ..density..), col = "black", fill = "#E69F00", bins = 15) +
  stat_function(fun = dnorm, args = list(mean = mean(hd$pwr_ldl),
                                         sd = sd(hd$pwr_ldl)),
                col = "#009E73", size = 2) +
  ggtitle(label = "Low Density Lipoprotein (ldl) - Power Transformation (power = 0.1)") +
  xlab("Low Density Lipoprotein (ldl)") +
  ylab("Density") +
  theme_minimal() +
  theme(plot.background = element_rect(colour = "black", fill=NA, size=1))

plt_pwrldl

plt_pwrobesity <- ggplot(data = hd, aes(x = pwr_obesity)) +
  geom_histogram(aes(y = ..density..), col = "black", fill = "#E69F00", bins = 15) +
  stat_function(fun = dnorm, args = list(mean = mean(hd$pwr_obesity),
                                         sd = sd(hd$pwr_obesity)),
                col = "#009E73", size = 2) +
  ggtitle(label = "Obesity - Power Transformation (power = -0.4)") +
  xlab("Obesity") +
  ylab("Density") +
  theme_minimal() +
  theme(plot.background = element_rect(colour = "black", fill=NA, size=1))

plt_pwrobesity

plt_pwralcohol <- ggplot(data = hd, aes(x = pwr_alcohol)) +
  geom_histogram(aes(y = ..density..), col = "black", fill = "#E69F00", bins = 15) +
  stat_function(fun = dnorm, args = list(mean = mean(hd$pwr_alcohol),
                                         sd = sd(hd$pwr_alcohol)),
                col = "#009E73", size = 2) +
  ggtitle(label = "Alcohol - Power Transformation (power = 0.4)") +
  xlab("Alcohol") +
```

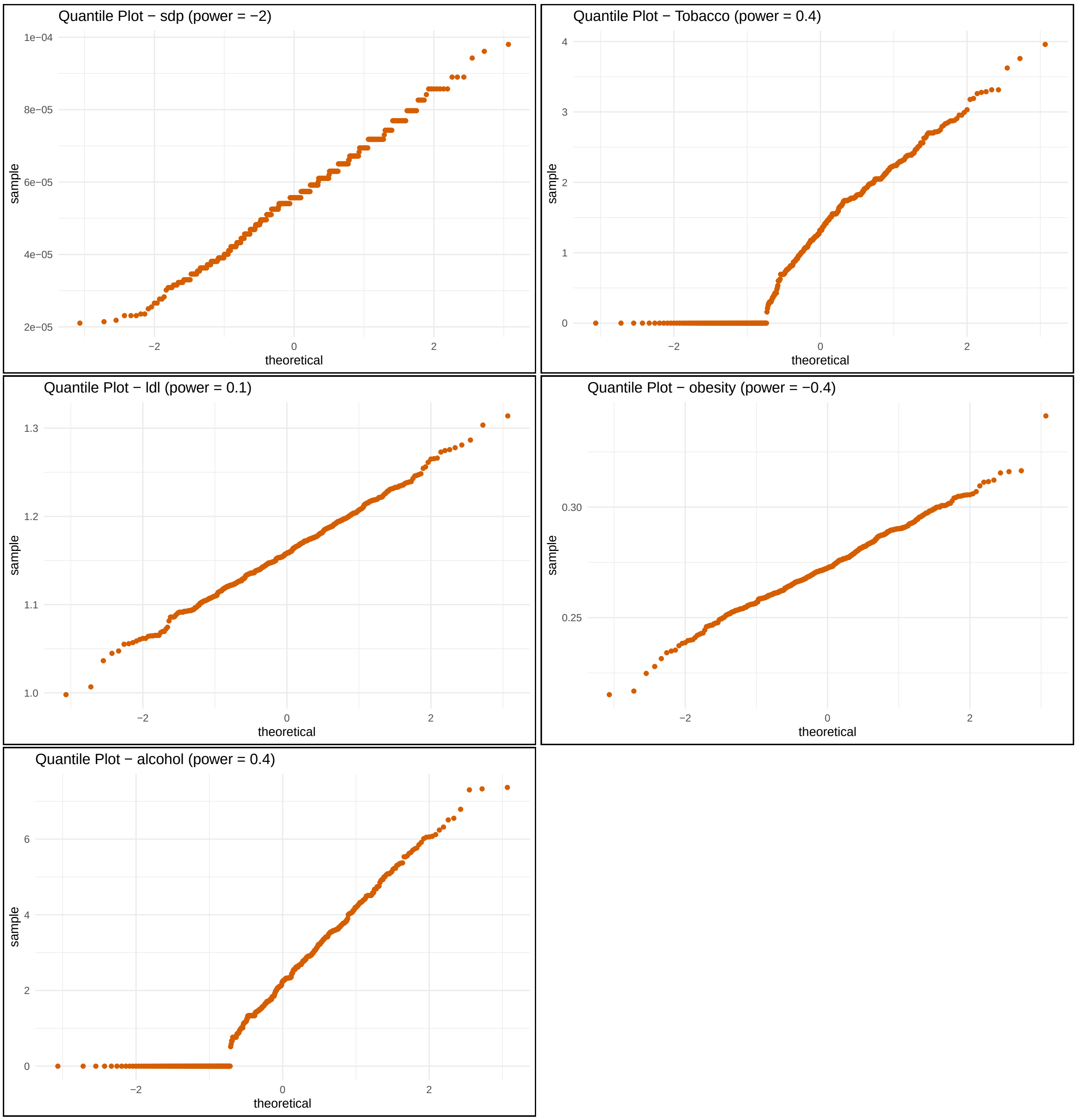
```
  ylab("Density") +
  theme_minimal() +
  theme(plot.background = element_rect(colour = "black", fill=NA, size=1))

plt_pwralcohol
```



Eight - Power Transformation Quantile Plots

```
qplt_pwrsdp <- ggplot(data = hd, aes(sample = pwr_sbp)) +
  stat_qq(col = "#D55E00") +
  ggtitle("Quantile Plot - sdp (power = -2)") +
  theme_minimal() +
  theme(plot.background = element_rect(colour = "black", fill=NA, size=1))

qplt_pwrsdp

qplt_pwrtobacco <- ggplot(data = hd, aes(sample = pwr_tobacco)) +
  stat_qq(col = "#D55E00") +
  ggtitle("Quantile Plot - Tobacco (power = 0.4)") +
  theme_minimal() +
  theme(plot.background = element_rect(colour = "black", fill=NA, size=1))

qplt_pwrtobacco

qplt_pwrldl <- ggplot(data = hd, aes(sample = pwr_ldl)) +
  stat_qq(col = "#D55E00") +
  ggtitle("Quantile Plot - ldl (power = 0.1)") +
  theme_minimal() +
  theme(plot.background = element_rect(colour = "black", fill=NA, size=1))

qplt_pwrldl

qplt_pwrobesity <- ggplot(data = hd, aes(sample = pwr_obesity)) +
  stat_qq(col = "#D55E00") +
  ggtitle("Quantile Plot - obesity (power = -0.4)") +
  theme_minimal() +
  theme(plot.background = element_rect(colour = "black", fill=NA, size=1))

qplt_pwrobesity
```

```r
qplt_pwralcohol <- ggplot(data = hd, aes(sample = pwr_alcohol)) +
  stat_qq(col = "#D55E00") +
  ggtitle("Quantile Plot - alcohol (power = 0.4)") +
  theme_minimal() +
  theme(plot.background = element_rect(colour = "black", fill=NA, size=1))

qplt_pwralcohol
```



Nine - Logistic Regression with Transformed Variables

```
logit_model_transformed <- glm(formula = chd ~ adiposity + typea + age +
                                 pwr_sbp + pwr_tobacco + pwr_ldl +
                                 pwr_obesity + pwr_alcohol,
                               family = binomial(link = "logit"), data = hd)

summary(logit_model_transformed)
```

```
##
## Call:
## glm(formula = chd ~ adiposity + typea + age + pwr_sbp + pwr_tobacco +
##     pwr_ldl + pwr_obesity + pwr_alcohol, family = binomial(link = "logit"),
##     data = hd)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7977  -0.8452  -0.4429   0.9735   2.5003
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.173e+01  4.905e+00  -4.430 9.43e-06 ***
## adiposity    3.193e-02  2.898e-02   1.102 0.270652
## typea        4.022e-02  1.205e-02   3.337 0.000847 ***
## age          4.712e-02  1.207e-02   3.904 9.47e-05 ***
## pwr_sbp     -4.317e+03  8.381e+03  -0.515 0.606477
## pwr_tobacco  4.194e-01  1.384e-01   3.031 0.002440 **
## pwr_ldl      8.489e+00  2.686e+00   3.161 0.001574 **
## pwr_obesity  2.096e+01  1.102e+01   1.902 0.057220 .
## pwr_alcohol  1.073e-02  6.358e-02   0.169 0.865998
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 596.11  on 461  degrees of freedom
## Residual deviance: 485.61  on 453  degrees of freedom
## AIC: 503.61
##
## Number of Fisher Scoring iterations: 5
```

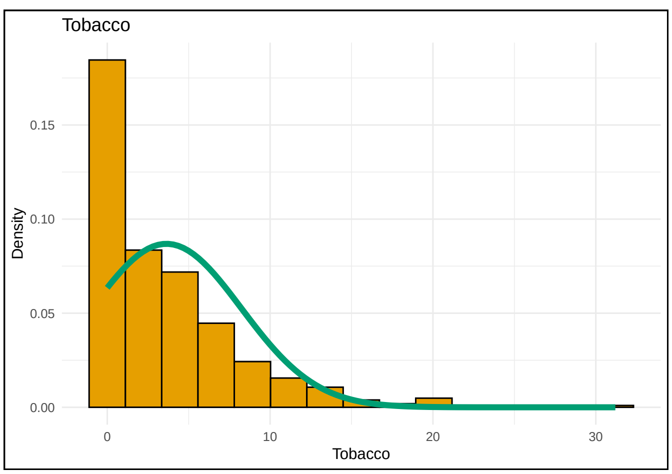
Problem 2 - Reporting

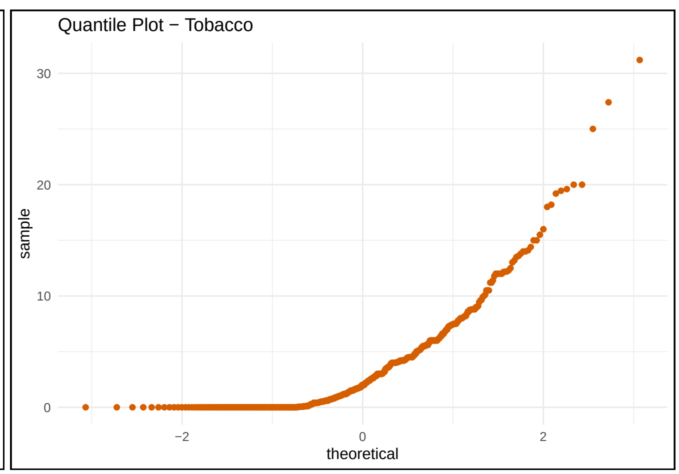
One - Univariate Table

Table 2: Univariate Statistics for Heart Disease Data

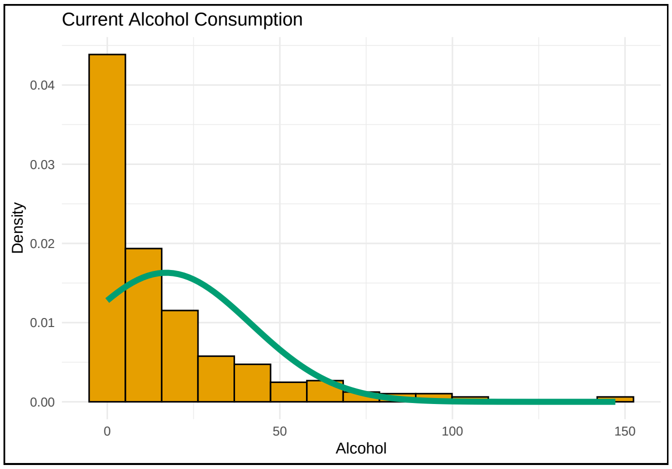
	mean	median	skew
sbp	138.327	134.000	1.173
tobacco	3.636	2.000	2.066
ldl	4.740	4.340	1.305
adiposity	25.407	26.115	-0.213
typea	53.104	53.000	-0.344
obesity	26.044	25.805	0.899
alcohol	17.044	7.510	2.298
age	42.816	45.000	-0.379

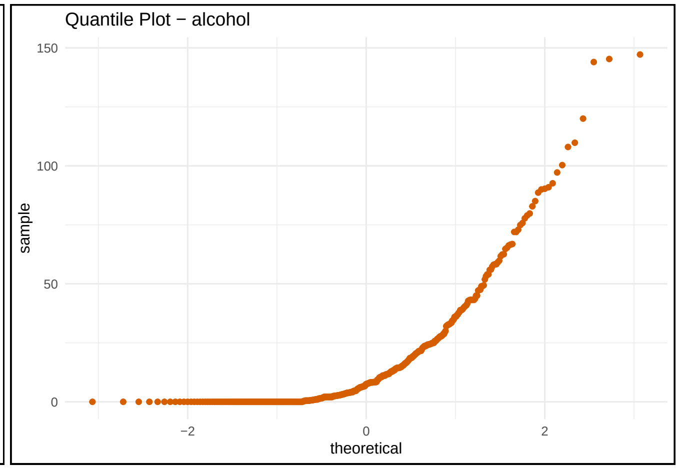
Two - Histogram and Quantile Plot of Tobacco





Three - Histogram and Quantile Plot of Alcohol



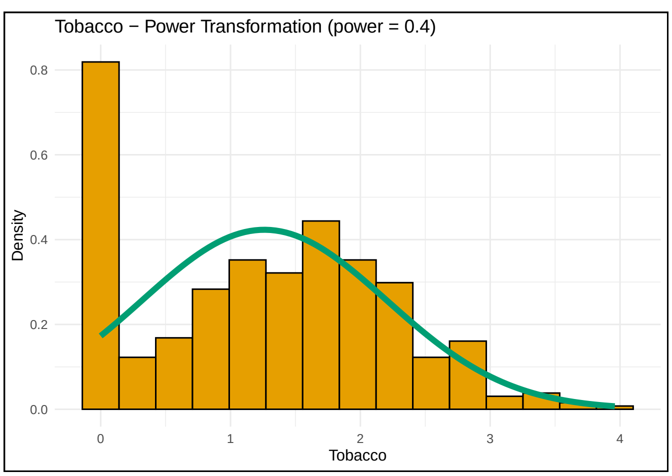


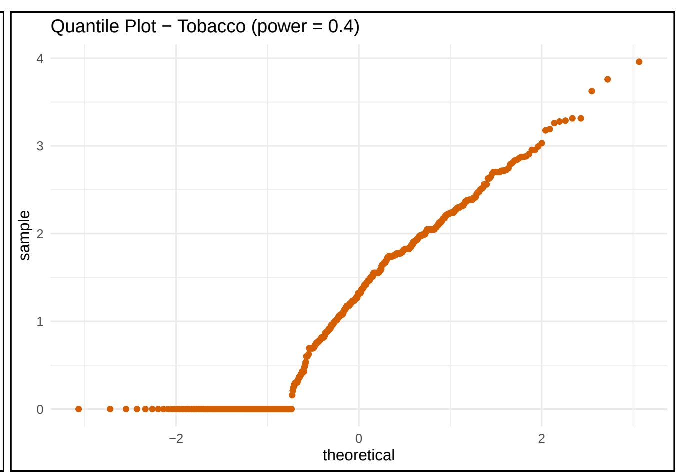
Four - Univariate Table for Power Transformation

Table 3: Univariate Statistics for Heart Disease Data

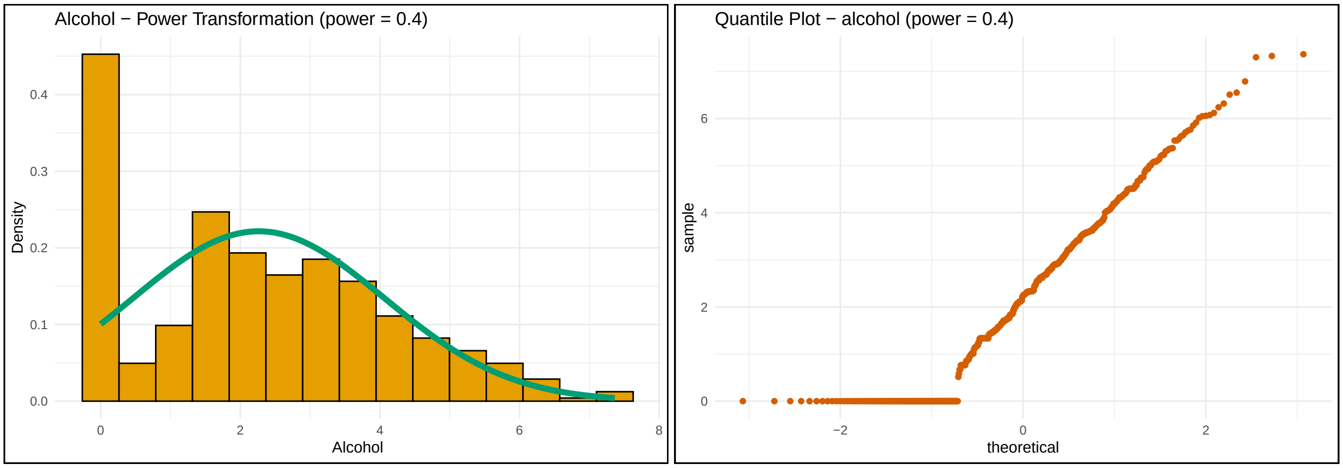
	mean	median	skew
pwr_alcohol	2.264	2.240	0.398
pwr_ldl	1.159	1.158	0.029
pwr_obesity	0.273	0.272	0.000
pwr_sbp	0.000	0.000	0.063
pwr_tobacco	1.261	1.320	0.134

Five - Histogram and Quantile Plot of Tobacco (Power Transformation)





Six - Histogram and Quantile Plot of Alcohol (Power Transformation)



Seven - Confidence Interval (Tobacco/Alcohol)

Eight - Confidence Interval (Alcohol)

Nine - Model Performance (c-statistics)

Textbook Questions

Question 2

Question 2

$$p_k(x) = \frac{\pi_k \left(1/\sqrt{2\pi}\,\sigma\right) e^{-(1/2\sigma^2)(x-\mu_k)^2}}{\sum_{l=1}^{K} \pi_l \left(1/\sqrt{2\pi}\,\sigma\right) e^{-(1/2\sigma^2)(x-\mu_l)^2}}$$

$$= \frac{\pi_k e^{-(1/2\sigma^2)(x-\mu_k)^2}}{\sum_{l=1}^{K} \pi_l e^{-(1/2\sigma^2)(x-\mu_l)^2}}$$

$$= \log p_k(x) = \log \pi_k - (1/2\sigma^2)(x-\mu_k)^2 - \log \sum_{l=1}^{K} \pi_l e^{-(1/2\sigma^2)(x-\mu_l)^2}$$

$$= \log \pi_k - (1/2\sigma^2)(x-\mu_k)^2 = \log \pi_k - \frac{1}{2\sigma^2} x^2 + \frac{\mu_k}{\sigma^2} x - \frac{\mu_k^2}{2\sigma^2}$$

$$= \delta_k(x) = \frac{\mu_k}{\sigma^2} x - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k$$

Question 6

(a) Givens: $\hat{\beta}_0 = -6,\ \hat{\beta}_1 = 0.05,\ X_1 = 40$ hrs, $X_2 = 3.5$

$$\begin{aligned}\hat{p}(X) &= \frac{e^{-6+0.05X_1+X_2}}{1+e^{-6+0.05X_1+X_2}} \\ &= \frac{e^{-6+0.05(40)+X_2}}{1+e^{-6+0.05(40)+X_2}} \\ &= \frac{e^{-0.5}}{1+e^{-0.5}} \\ &= 0.37754\end{aligned}$$

(b) The student in part (a) needs to study 50 hours in order to have a 50% chance of getting an A in the class.

Givens: $\hat{p}(X) = 0.5,\ X_2 = 3.5$

$$\begin{aligned}\hat{p}(X) &= \frac{e^{-6+0.05X_1+X_2}}{1+e^{-6+0.05X_1+X_2}} \\ 0.5 &= \frac{e^{-6+0.05X_1+3.5}}{1+e^{-6+0.05X_1+3.5}} \\ 0.5(1+e^{-6+0.05X_1+3.5}) &= e^{-6+0.05X_1+3.5} \\ 0.5+0.5e^{-2.5+0.05X_1} &= e^{-2.5+0.05X_1} \\ 0.5 &= 0.5e^{-2.5+0.05X_1} \\ 1 &= e^{-2.5+0.05X_1} \\ log(1) &= log(e^{-2.5+0.05X_1}) \\ 0 &= -2.5+0.05X_1 \\ 2.5 &= 0.05X_1 \\ X_1 &= 50\end{aligned}$$

Question 7

$$p_1(4) = \frac{0.8\,e^{-(1/72)(4-10)^2}}{0.8\,e^{-(1/72)(4-10)^2} + 0.2\,e^{-(1/72)(4-0)^2}}$$

$$= \frac{0.8\,e^{-0.5}}{0.8\,e^{-0.5} + 0.2\,e^{0.222\ldots}}$$

$$= 0.75185\ldots$$

$$\approx 0.752 \qquad \blacksquare$$

Question 8

Since we are unsure what the true test data error rate is for the K Nearest neighbor we should not use this method. It could be that the error rate for the test data was 30% while the training data is 6%, but without knowing for sure it would be unwise to use this method. Additionally, if you multiply the average error rate by 2, then you get 36% which is greater than the 30% for logistic regression. With this being said it would be best to use logistic regression.

Question 10

Question 12