EDA in Practice: Heart Disease

Import the dataset, heartdisease.csv

Find the structure and data types of the variables.

Missing values

Plot the missing values using the plot_missing() function.

```

```

Transform variables

Create a data object, hd2. a. Use mutate() to create new variables, “sex”, “fbs”, “exang”, “cp”, and “restecg” and using the if_else function. b. Create factors of “slope”, “ca”, “thal”

```
hd2 <- %>%
  mutate(sex = if_else(""),
         fbs = if_else(fbs == 1, ">120", "<=120"),
         exang = if_else(
         cp =  if_else(cp == 1, "ATYPICAL ANGINA",
               if_else(cp == 2, "NON-ANGINAL PAIN", "ASYMPTOMATIC")),
         restecg = if_else(restecg == 0, "NORMAL",
               if_else(restecg == 1, "ABNORMALITY", "PROBABLE OR DEFINITE")),
         slope =
         ca =
         thal =
         target = if_else(target == 1, "YES", "NO")
         ) %>%
  mutate_if(is.character, as.factor) %>%
  dplyr::select(target, sex, fbs, exang, cp, restecg, slope, ca, thal, everything
())
```

Summary statistics

Find the summary statistics of the dataset, hd2.

```

```

Using the dataframe, hd2, create boxplots of columns “age”:”oldpeak” using base graphics.

Data Visualization

Create a barplot for the target variable, Heart Disease. Add an x-axis label, “Heart Disease” and y-axis label, “Count”.

```
ggplot(data2, aes(, fill=data2$target)) +
  geom_bar() +
  +
  +
```

```
  ggtitle("Analysis of Presence and Absence of Heart Disease") +
  scale_fill_discrete(name = "Heart Disease", labels = c("Absence", "Presence"))
```

Find the proportion (frequency) of the dataframe, hd2, where target equals “NO” and target equals “YES”. Use the $ syntax.

Group by age, c*ount, n,* the frequency of the “age” variable. Filter values where count > 10.

Add a main plot title, x-axis label, “age” and y-axis, “AgeCount”.

```
data2 %>%
  %>%
  %>%
  %>%
  ggplot()+
  geom_col(aes(), fill = "navyblue")+
  ggtitle("Age Analysis") +
  +
```

Create a bar graph of the variable, “cp” representing levels of the target variable. Fill in the blanks.

```
ggplot(data2, aes(, fill =  ))+
  geom_bar(position = "fill")+
  ggtitle("cp")
```

Let’s look at the distribution of Male and Female population across the “age” parameter. Use the histogram geometry.

```
data2 %>%
  ggplot(aes( ))+
  ()+
  xlab("Age") +
  ylab("Number")+
  guides(fill = guide_legend(title = "Gender"))
```

Let’s look at how cholesterol levels are represented in the data. Using the input dataframe, hd2, create a scatterplot of “age” and “chol”. Use options color to color the points by sex and size to represent the size of the point by the cholesterol level.

```
  %>%
  ggplot(aes(x=,y=,color=, size=))+
  geom_ (alpha=0.7)+
  xlab() +
  ylab() +
  guides(fill = guide_legend(title = "Gender"))
```

Let's compare blood pressure across pain type. Using the input dataframe, hd2, create multiple boxplots of "trestbps" by "sex". Lastly use facet_grid() with the "cp" variable.

```
  %>%
  ggplot(aes(x=,y=))+
  geom_ (fill="darkorange")+
  xlab()+
  ylab()+
  facet_grid(~cp)
```

Let's compare cholestrol across pain type. Create multiple boxplots of "chol" by "sex". *Fill* the box using a hex code. Lastly use facet_grid() with the "cp" variable.

```
data2 %>%
  ggplot(aes(x=,y=))+
  geom_ (fill="#D55E00")+
  xlab("")+
  ylab("")+
  facet_grid(~cp)
```

Find the correlations of columns 10:14. Map the correlations to an object "corr_heart". Use the dataframe, corr_heart, to create a corrplot, keeping only the upper diagonal. Use method=ellipse.

```
cor_heart <- cor()
cor_heart

corrplot(cor_heart, method = "ellipse", type="upper",)
```

Option #2:

```
ggcorrplot(cor_heart,lab = T)
ggcorr(cor_heart, label = T, label_round = 2)
```

Use the library, gridExtra, to create multiple bar plots of "sex", "fbs", "exang". Use the data frame, hd2, as the input data frame.

```
library(gridExtra)

grid.arrange(
  ggplot(data2, aes(x = sex, fill = target))+
  geom_bar(position = "fill"),

  ggplot(data2, aes(x = fbs, fill = target))+
  geom_bar(position = "fill"),

  ggplot(data2, aes(x = exang, fill = target))+
  geom_bar(position = "fill"), nrow = 3  )
```

Create scatterplots using “age” and “trestbps”. Use the options color() and shape() to color code and size the points of the “target” variable. Include a best fit line.

Create scatterplots using “age” and “chol”. Use the options color() and shape() to color code and size the points of the “target” variable. Include a best fit line.

Create scatterplots using “age” and “thalach”. Use the options color() and shape() to color code and size the points of the “target” variable. Include a best fit line.

Create scatterplots using “age” and “oldpeak”. Use the options color() and shape() to color code and size the points of the “target” variable. Include a best fit line.