```
#Activity 1 -Basic Imputation Methods
churn <- Churn_Train
summary(churn)
summary(is.na(churn))

#Replace missing value with mean
churn$Total_Charges[is.na(churn$Total_Charges)]<-mean(churn$Total_Charges,na.rm=TRUE)

#Replace missing value with median
churn$Total_Charges[is.na(churn$Total_Charges)]<-median(churn$Total_Charges,na.rm=TRUE)

churn$Total_Charges
library(ggplot2)
library(dplyr)
library(cowplot)

ggplot(churn,aes(Total_Charges))+
  geom_histogram(color="#000000",fill="#0099F8")+
  ggtitle("Variable distribution")+
  theme_classic()+
  theme(plot.title=element_text(size=18))

totalc_mode <- table(Churn_Train$Total_Charges)
mode_totalc <- as.numeric(names(totalc_mode[totalc_mode == max(totalc_mode)]))

#Perform simple value imputation and view the data
value_imputed<-data.frame(
  original=churn$Total_Charges,
  imputed_zero=replace(churn$Total_Charges,is.na(churn$Total_Charges),0),
  imputed_mean=replace(churn$Total_Charges,is.na(churn$Total_Charges),mean(churn$Total_Charges,na.rm=TRUE)),

imputed_median=replace(churn$Total_Charges,is.na(churn$Total_Charges),median(churn$Total_Charges,na.rm=TRUE)),
  imputed_mode=replace(churn$Tenure,is.na(churn$Total_Charges),mode_totalc)
)
value_imputed

#Create histograms after imputation
h1 <- ggplot(value_imputed, aes(x = original)) +
  geom_histogram(fill = "#ad1538", color = "#000000", position =
                   "identity") +
  ggtitle("Original distribution") +
  theme_classic()
h2 <- ggplot(value_imputed, aes(x = imputed_zero)) +
  geom_histogram(fill = "#15ad4f", color = "#000000", position =
                   "identity") +
  ggtitle("Zero-imputed distribution") +
  theme_classic()
h3 <- ggplot(value_imputed, aes(x = imputed_mean)) +
  geom_histogram(fill = "#1543ad", color = "#000000", position =
                   "identity") +
  ggtitle("Mean-imputed distribution") +
  theme_classic()
h4 <- ggplot(value_imputed, aes(x = imputed_median)) +
  geom_histogram(fill = "#ad8415", color = "#000000", position =
                   "identity") +
  ggtitle("Median-imputed distribution") +
  theme_classic()
plot_grid(h1, h2, h3, h4, nrow = 2, ncol = 2)

#Activity 2 - Impute Missing Values with MICE
library(mice)

churn_numeric <- Churn_Train %>%
  select(Monthly_Charges, Total_Charges)

# Check the missing data pattern
md.pattern(churn_numeric)

# Perform MICE imputation methods
mice_imputed <- data.frame(
  original = Churn_Train$Total_Charges,
  imputed_pmm = complete(mice(churn_numeric, method = "pmm"))$Total_Charges,
  imputed_cart = complete(mice(churn_numeric, method = "cart"))$Total_Charges,
  imputed_lasso = complete(mice(churn_numeric, method = "lasso.norm"))$Total_Charges
)
mice_imputed
```

```
h1 <- ggplot(mice_imputed, aes(x = original)) +
  geom_histogram(fill = "#ad1538", color = "#000000", position =
                   "identity") +
  ggtitle("Original distribution") +
  theme_classic()
h2 <- ggplot(mice_imputed, aes(x = imputed_pmm)) +
  geom_histogram(fill = "#15ad4f", color = "#000000", position =
                   "identity") +
  ggtitle("Pmm-imputed distribution") +
  theme_classic()
h3 <- ggplot(mice_imputed, aes(x = imputed_cart)) +
  geom_histogram(fill = "#1543ad", color = "#000000", position =
                   "identity") +
  ggtitle("Cart-imputed distribution") +
  theme_classic()
h4 <- ggplot(mice_imputed, aes(x = imputed_lasso)) +
  geom_histogram(fill = "#ad8415", color = "#000000", position =
                   "identity") +
  ggtitle("Lasso-imputed distribution") +
  theme_classic()
plot_grid(h1, h2, h3, h4, nrow = 2, ncol = 2)


#Activity 3 - Imputation with R missForest Package
churn_numeric <- Churn_Train %>%
  select(Monthly_Charges, Total_Charges)

sum(is.na(Churn_Train))
library(missForest)
missForest_imputed<-data.frame(
  original=churn_numeric$Total_Charges,
  imputed_missForest=missForest(churn_numeric)$ximp$Total_Charges
)
missForest_imputed

h1 <- ggplot(missForest_imputed, aes(x = original)) +
  geom_histogram(fill = "#ad1538", color = "#000000", position =
                   "identity") +
  ggtitle("Original distribution") +
  theme_classic()

h2 <- ggplot(missForest_imputed, aes(x = imputed_missForest)) +
  geom_histogram(fill = "#15ad4f", color = "#000000", position =
                   "identity") +
  ggtitle("missForest-imputed distribution") +
  theme_classic()
plot_grid(h1, h2, nrow = 1, ncol = 2)


#Activity 4: Normalize data with scaling methods
#1 Log Transformation
log_scale = log(as.data.frame(Churn_Train$Total_Charges))

#2 Min-Max Scaling
library(caret)
process <- preProcess(as.data.frame(Churn_Train$Total_Charges),
                      method=c("range"))
norm_scale <- predict(process, as.data.frame(Churn_Train$Total_Charges))

#3 Standard scaling
scale_data <- as.data.frame(scale(Churn_Train$Total_Charges))


#Activity 5: Feature Encoding
#1 Label Encoding
gender_encode <- ifelse(Churn_Train$Gender == "male",1,0)
table(gender_encode)

embarked_encode <- ifelse(Churn_Train$`Multiple Lines` == "Yes",1,
                          ifelse(Churn_Train$`Multiple Lines` == "No",2,
                                 ifelse(Churn_Train$`Multiple Lines` == "No phone service",3,0)))
table(embarked_encode)

#2 One-Hot Encoding
new_dat = data.frame(Churn_Train$Total_Charges,Churn_Train$Gender,Churn_Train$`Multiple Lines`)
summary(new_dat)

library(caret)
```

```
dmy <- dummyVars(" ~ .", data = new_dat, fullRank = T)
dat_transformed <- data.frame(predict(dmy, newdata = new_dat))
glimpse(dat_transformed)

#3 Encoding Continuous (or Numeric) Variables
summary(new_dat$Churn_Train.Total_Charges)

bins <- c(-Inf, 399.3, 3786.6, Inf)
bin_names <- c("Low", "Mid50", "High")
new_dat$new_TotalCharges <- cut(new_dat$Churn_Train.Total_Charges, breaks =
                          bins, labels = bin_names)
summary(new_dat$Churn_Train.Total_Charges)
summary(new_dat$new_TotalCharges)

churn %>%
  eda_web_report(target = "Total_Charges", subtitle = "lab homework",
                 output_dir = "C:/Users/jing2/OneDrive/Desktop/Data Science", output_file = "EDA.html", theme
= "blue")

rmarkdown::render("Lab3.Rmd", output_dir = "C:/Users/jing2/OneDrive/Desktop/Data Science")
```