````
---
title: "sta302 final project"
output:
  pdf_document:
    latex_engine: xelatex
date: "2023-08-13"
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

```{r setup, include=FALSE}
library(tidyr)
library(dplyr)
library(car)
library(ggplot2)
library(MASS)
```


```{r}
Read_data = read.csv("adm_data.csv", header = TRUE)
```

# Exploratory data analysis section

##Summary data
```{r}
summ_result <- summary(Read_data[, !(names(Read_data) %in% "Serial.No.")])
print(summ_result)
```

## Make Summary data more clearly
```{r}
numeric_data <- Read_data[, !(names(Read_data) %in% "Serial.No.")]
numeric_data <- numeric_data[sapply(numeric_data, is.numeric)]
result_table <- data.frame(
  Min = sapply(numeric_data, min, na.rm = TRUE),
  `1st Qu.` = sapply(numeric_data, function(x) quantile(x, 0.25, na.rm = TRUE)),
  Median = sapply(numeric_data, median, na.rm = TRUE),
  Mean = sapply(numeric_data, mean, na.rm = TRUE),
  `3rd Qu.` = sapply(numeric_data,function(x) quantile(x, 0.75, na.rm = TRUE)),
  Max = sapply(numeric_data, max, na.rm = TRUE),
  Std.Dev = sapply(numeric_data, sd, na.rm = TRUE)
)
rownames(result_table) <- names(numeric_data)
print(result_table)
```

## Histogram
```{r}
hist(Read_data$Chance.of.Admit, breaks = 60, main="Original Data", xlab="Chance of Admit")
```


# Scatter plot
```{r}
plot1 <- ggplot(Read_data, aes(x=GRE.Score, y=Chance.of.Admit)) +
  geom_point() +
  geom_smooth(se=FALSE, method="lm") +
  ggtitle("Scatter plot of Chance.of.Admit with GRE.Score")

plot2 <- ggplot(Read_data, aes(x=TOEFL.Score, y=Chance.of.Admit)) +
````
````

```
  geom_point() +
  geom_smooth(se=FALSE, method="lm") +
  ggtitle("Scatter plot of Chance.of.Admit with TOEFL.Score")

plot3 <- ggplot(Read_data, aes(x=CGPA, y=Chance.of.Admit)) +
  geom_point() +
  geom_smooth(se=FALSE, method="lm") +
  ggtitle("Scatter plot of Chance.of.Admit with CGPA")

grid.arrange(plot1, plot2, plot3,ncol=1)
```

```
# Side_by_Side plot
```{r}
ggplot(data=Read_data, aes(x=factor(Research), y=Chance.of.Admit)) +
  geom_boxplot(color='green', fill='blue') +
  labs(title="Research with Chance.of.Admit",
       x="Research",
       y="Chance of Admit") +
  theme_minimal()
```

# Model Development

## Varibale of transformation
```{r}
layout(matrix(1:4, 2, 2))

hist(Read_data$Chance.of.Admit, breaks = 60, main="Original Data", xlab="Chance of Admit")

hist(log(Read_data$Chance.of.Admit), breaks = 60, main="Log-transformed Histogram",
xlab="log(Chance of Admit)")

hist(sqrt(Read_data$Chance.of.Admit), breaks = 60, main="Square Root-transformed
Histogram", xlab="sqrt(Chance of Admit)")

squared_chance_of_admit <- Read_data$Chance.of.Admit^2
hist(squared_chance_of_admit, breaks = 60, main="Squared Data", xlab="Squared Chance of
Admit")
```



## Check Multicollineaity
```{r}
full_model <- lm(Chance.of.Admit ~ . - Serial.No., data=Read_data)
```

```{r}
vif(full_model)
```



```{r}
Non_Multicollineaity_full_model <- lm(Chance.of.Admit ~ . - Serial.No. - Research - CGPA,
data = Read_data)
vif(Non_Multicollineaity_full_model)
```

```{r}
auto_reduced_model <- stepAIC(Non_Multicollineaity_full_model, direction="both")
```

```{r}
summary(auto_reduced_model)
```
```

```
```

```{r}
summary(Non_Multicollineaity_full_model)$adj.r.squared
```

```{r}
summary(auto_reduced_model)$adj.r.squared
```

```{r}
AIC(Non_Multicollineaity_full_model)
```

```{r}
AIC(auto_reduced_model)
```
```{r}
BIC(Non_Multicollineaity_full_model)
```

```{r}
BIC(auto_reduced_model)
```

```{r}
adj_r2_full <- summary(Non_Multicollineaity_full_model)$adj.r.squared
aic_full <- AIC(Non_Multicollineaity_full_model)
bic_full <- BIC(Non_Multicollineaity_full_model)

adj_r2_reduced <- summary(model_auto_reduced)$adj.r.squared
aic_reduced <- AIC(auto_reduced_model)
bic_reduced <- BIC(auto_reduced_model)

results_table <- data.frame(
  Model = c("Non_Multicollineaity_full_model", "model_auto_reduced"),
  Adjusted_R_Squared = c(adj_r2_full, adj_r2_reduced),
  AIC = c(aic_full, aic_reduced),
  BIC = c(bic_full, bic_reduced)
)
results_table

```

```{r}
anova(auto_reduced_model,Non_Multicollineaity_full_model )
```

# Residual Plot
```{r}
res <- auto_reduced_model$residuals
y_hat <- fitted(auto_reduced_model)
plot(y_hat, res)
```

## Residual vs Predictors
```{r}
par(mfrow = c(1,4))
plot(Read_data$GRE.Score,res)
plot(Read_data$TOEFL.Score,res)
plot(Read_data$University.Rating,res)
plot(Read_data$LOR ,res)
```

```r
layout(matrix (c(1,2,3,4),2,2))
plot(model_auto_reduced)
```

## varibale transformation

```r
Read_data$Transformed_Chance_of_Admit <-  Read_data$Chance.of.Admit^2
full_model <- lm(Transformed_Chance_of_Admit ~ . - Serial.No., data=Read_data)

```

```r
vif_values <- vif(full_model)
print(vif_values)
```

```r
full_model1 <- lm(Transformed_Chance_of_Admit ~ . - Serial.No. - Research - CGPA ,
data=Read_data)
vif_values <- vif(full_model1)
print(vif_values)
```

```r
model_auto_reduced1 <- stepAIC(full_model1, direction="both")
```

```r
summary(model_auto_reduced1)
```

```r
anova(model_auto_reduced1,full_model1)
```

```r
res1 <- model_auto_reduced1$residuals
y_hat <- fitted(model_auto_reduced1)
plot(y_hat, res1)
```

```r
layout(matrix (c(1,2,3,4),2,2))
plot(model_auto_reduced1)
```