Ejercicio # 2

Ronald Bailey

2023-05-28

Ejercicio #2:

Para este ejercicio se le solicita que desarrolle las siguientes actividades utilizando RStudio Con el dataset Admissions adjunto a este laboratorio realice lo siguiente:

```
dataset = read.csv("Admisions.csv")
```

1. Realice un análisis estadístico sobre todas las variables del dataset, recuerde que pude usar la función summary().

```
summary(dataset)
```

```
##    Serial.No.      GRE.Score      TOEFL.Score    University.Rating
##  Min.   :  1.0   Min.   :290.0   Min.   : 92.0   Min.   :1.000
##  1st Qu.:125.8   1st Qu.:308.0   1st Qu.:103.0   1st Qu.:2.000
##  Median :250.5   Median :317.0   Median :107.0   Median :3.000
##  Mean   :250.5   Mean   :316.5   Mean   :107.2   Mean   :3.114
##  3rd Qu.:375.2   3rd Qu.:325.0   3rd Qu.:112.0   3rd Qu.:4.000
##  Max.   :500.0   Max.   :340.0   Max.   :120.0   Max.   :5.000
##       SOP             LOR             CGPA         Research
##  Min.   :1.000   Min.   :1.000   Min.   :6.800   Min.   :0.00
##  1st Qu.:2.500   1st Qu.:3.000   1st Qu.:8.127   1st Qu.:0.00
##  Median :3.500   Median :3.500   Median :8.560   Median :1.00
##  Mean   :3.374   Mean   :3.484   Mean   :8.576   Mean   :0.56
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:9.040   3rd Qu.:1.00
##  Max.   :5.000   Max.   :5.000   Max.   :9.920   Max.   :1.00
##  Chance.of.Admit
##  Min.   :0.3400
##  1st Qu.:0.6300
##  Median :0.7200
##  Mean   :0.7217
##  3rd Qu.:0.8200
##  Max.   :0.9700
```

Según la descripción proporcionada por la función `summary()`, el conjunto de datos tiene varias variables:

- **Serial.No.** : Esta parece ser una identificación única para cada estudiante. Rango de 1 a 500, y parece estar distribuido uniformemente dado que la mediana y la media son aproximadamente la misma (250.5).

- **GRE.Score** : Este es el puntaje GRE de los estudiantes. La puntuación mínima es de 290, la máxima es de 340, y la media es de 316.5, lo que indica que los puntajes son relativamente altos en promedio.
- **TOEFL.Score** : Este es el puntaje TOEFL de los estudiantes. La puntuación mínima es de 92, la máxima es de 120 y la media es de 107.2, lo que también indica un nivel de puntajes relativamente alto.
- **University.Rating** : Esta es la calificación de la universidad. Las calificaciones van de 1 a 5, y la media está en 3.114, lo que sugiere que las calificaciones están distribuidas bastante uniformemente.
- **SOP** : Esta parece ser una medida de la calidad de la declaración de propósito de los estudiantes. Va de 1 a 5, y la media es 3.374, lo que indica una calidad promedio bastante buena.
- **LOR** : Esta es una medida de la calidad de las cartas de recomendación. También va de 1 a 5, y la media es 3.484, lo que también indica una buena calidad promedio.
- **CGPA** : Este es el promedio de calificaciones acumulado de los estudiantes. Va de 6.8 a 9.92, y la media es 8.576, lo que indica un alto nivel de rendimiento académico.
- **Research** : Este parece ser un indicador binario de si el estudiante ha hecho investigación o no. La media es 0.56, lo que indica que más de la mitad de los estudiantes ha hecho investigación.
- **Chance.of.Admit** : Esta es la probabilidad de admisión. Va de 0.34 a 0.97, y la media es 0.7217, lo que indica que la probabilidad promedio de admisión es bastante alta.

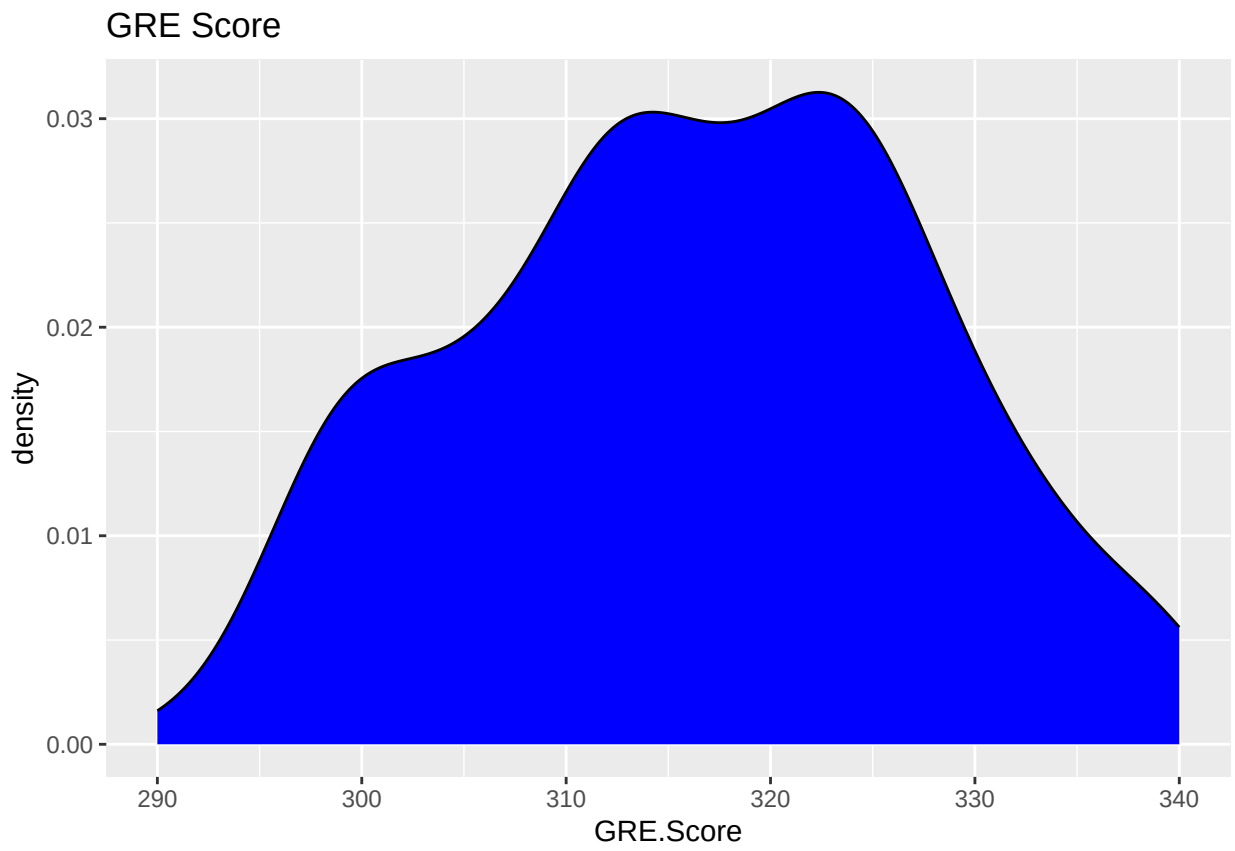
Cabe señalar que aunque estos análisis proporcionan información valiosa, es importante complementarlos con visualizaciones de datos y análisis adicionales, como verificar la correlación entre diferentes variables, realizar pruebas de hipótesis, etc.

2. Realice una gráfica de densidad para cada una de las variables numéricas en el dataset: GRE.Score, TOEFEL.Score, CGPA y Chance of Admit.

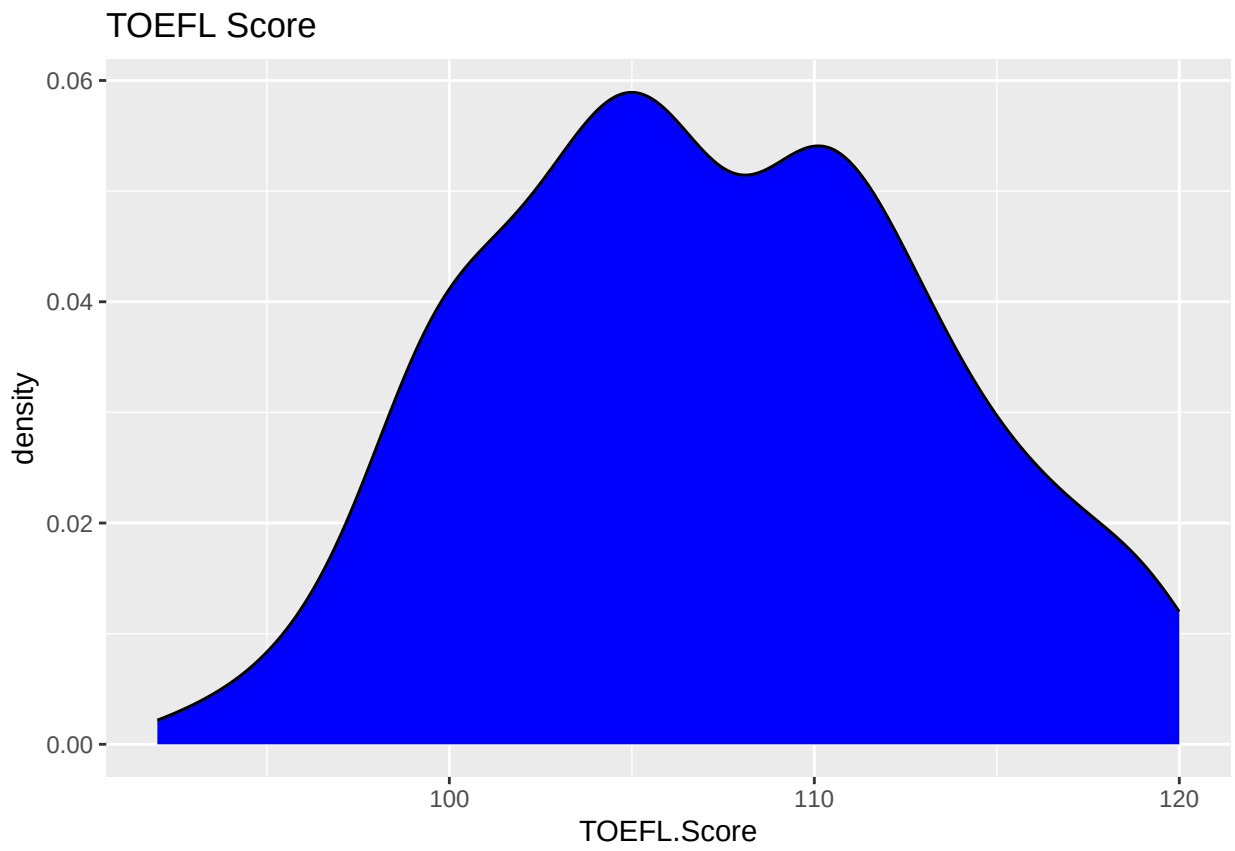
```
# Cargar la librería ggplot2
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

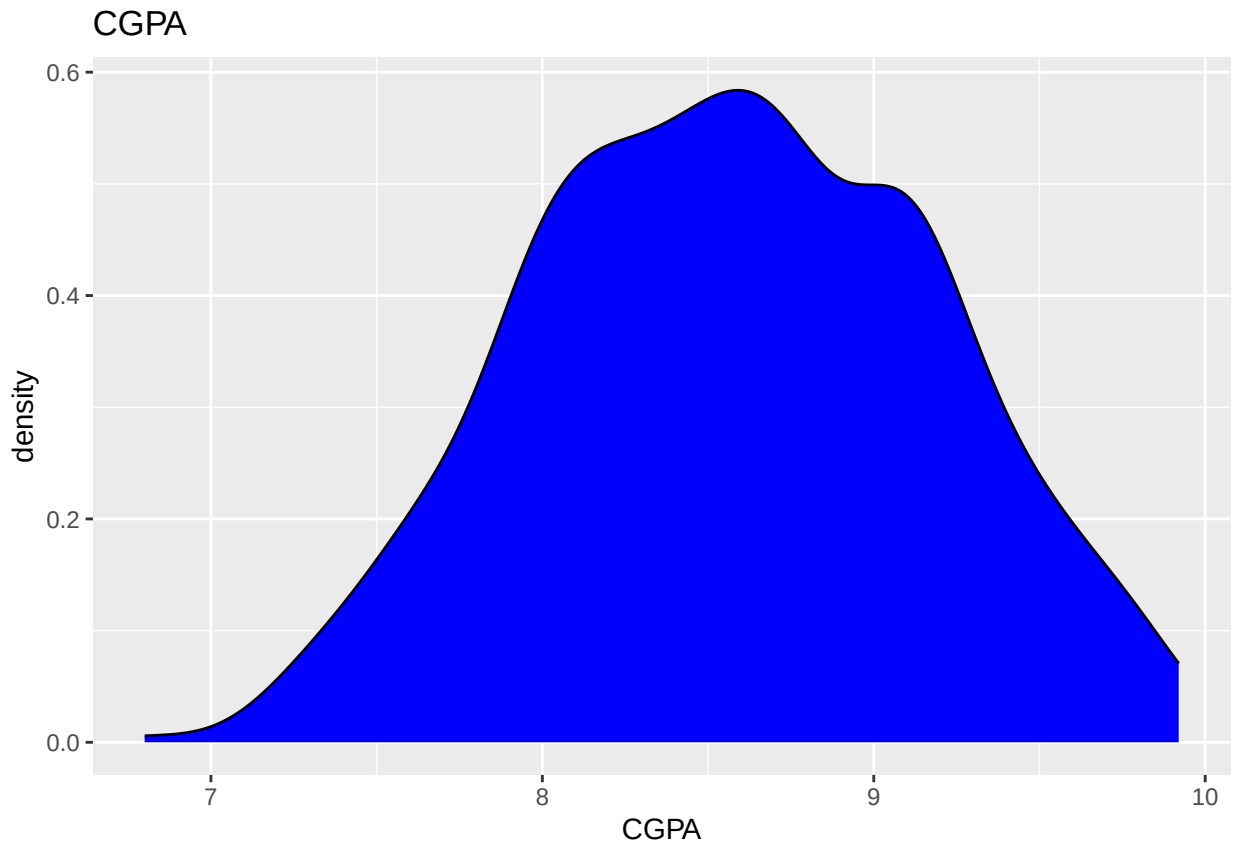
```
# Crear gráficos de densidad
ggplot(dataset, aes(x = GRE.Score)) + geom_density(fill = "blue") + labs(title = "GRE Score")
```



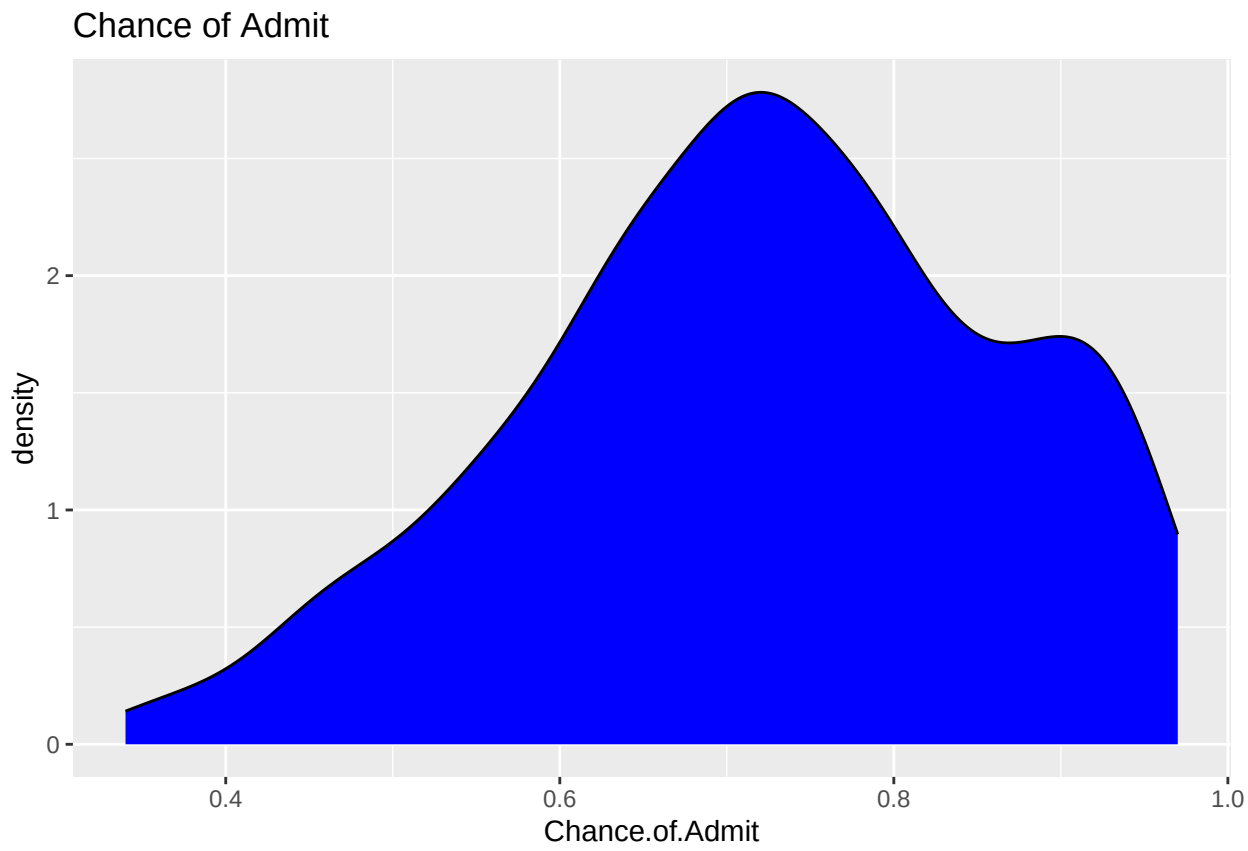

```
ggplot(dataset, aes(x = TOEFL.Score)) + geom_density(fill = "blue") + labs(title = "TOEFL Score")
```




```
ggplot(dataset, aes(x = CGPA)) + geom_density(fill = "blue") + labs(title = "CGPA")
```




```
ggplot(dataset, aes(x = Chance.of.Admit)) + geom_density(fill = "blue") + labs(title = "Chance of Admit
```




```
# Cargar la biblioteca corrplot
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.2.3
```

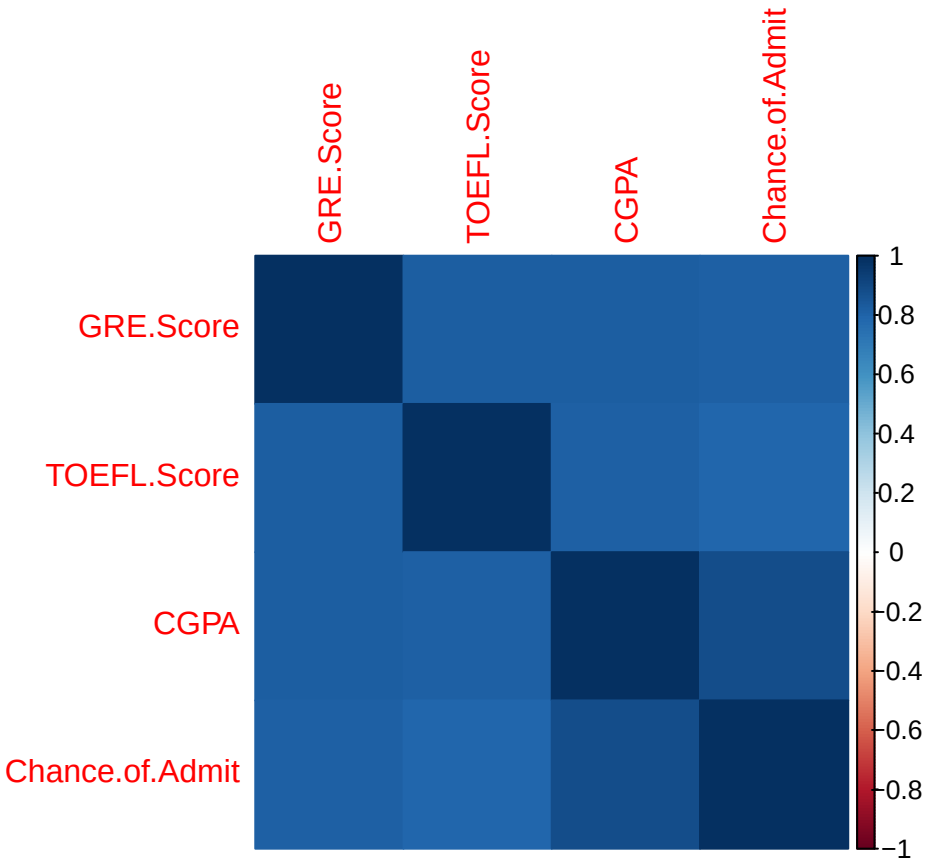
```
## corrplot 0.92 loaded
```

```
# Subconjunto de datos con las variables de interés
data_subset <- dataset[, c("GRE.Score", "TOEFL.Score", "CGPA", "Chance.of.Admit")]

# Calcular la matriz de correlación
cor_matrix <- cor(data_subset)
print(cor_matrix)
```

```
##                  GRE.Score TOEFL.Score      CGPA Chance.of.Admit
## GRE.Score        1.0000000   0.8272004 0.8258780       0.8103506
## TOEFL.Score      0.8272004   1.0000000 0.8105735       0.7922276
## CGPA             0.8258780   0.8105735 1.0000000       0.8824126
## Chance.of.Admit 0.8103506   0.7922276 0.8824126       1.0000000
```

```
# Crear el mapa de calor de correlación
corrplot(cor_matrix, method = "color")
```



4. Realice comentarios sobre el análisis estadístico de las variables numéricas y la gráfica de correlación.

El análisis estadístico de las variables numéricas revela información importante sobre el conjunto de datos:

- **GRE.Score**: Los puntajes de GRE van desde 290 hasta 340, con una media de 316.5. Esto indica que los puntajes son relativamente altos en promedio, lo que puede sugerir que los estudiantes tienen un buen desempeño en los exámenes GRE.
- **TOEFL.Score**: Los puntajes de TOEFL varían desde 92 hasta 120, con una media de 107.2. Al igual que los puntajes de GRE, los puntajes de TOEFL también son relativamente altos en promedio, lo que puede indicar un nivel de dominio del idioma inglés para los estudiantes.
- **CGPA**: Las calificaciones acumuladas promedio (CGPA) oscilan entre 6.8 y 9.92, con una media de 8.576. Esta media alta sugiere que los estudiantes tienen un buen desempeño académico en general.
- **Chance.of.Admit**: La probabilidad de admisión varía entre 0.34 y 0.97, con una media de 0.7217. La media relativamente alta indica que los estudiantes tienen una probabilidad promedio de admisión bastante alta.

En cuanto a la gráfica de correlación, se observa lo siguiente:

- Las variables `GRE.Score`, `TOEFL.Score` y `CGPA` están fuertemente correlacionadas entre sí y también tienen una correlación significativa con la variable `Chance.of.Admit`. Esto sugiere que los estudiantes con altos puntajes en el GRE, TOEFL y calificaciones acumuladas tienen una mayor probabilidad de admisión.

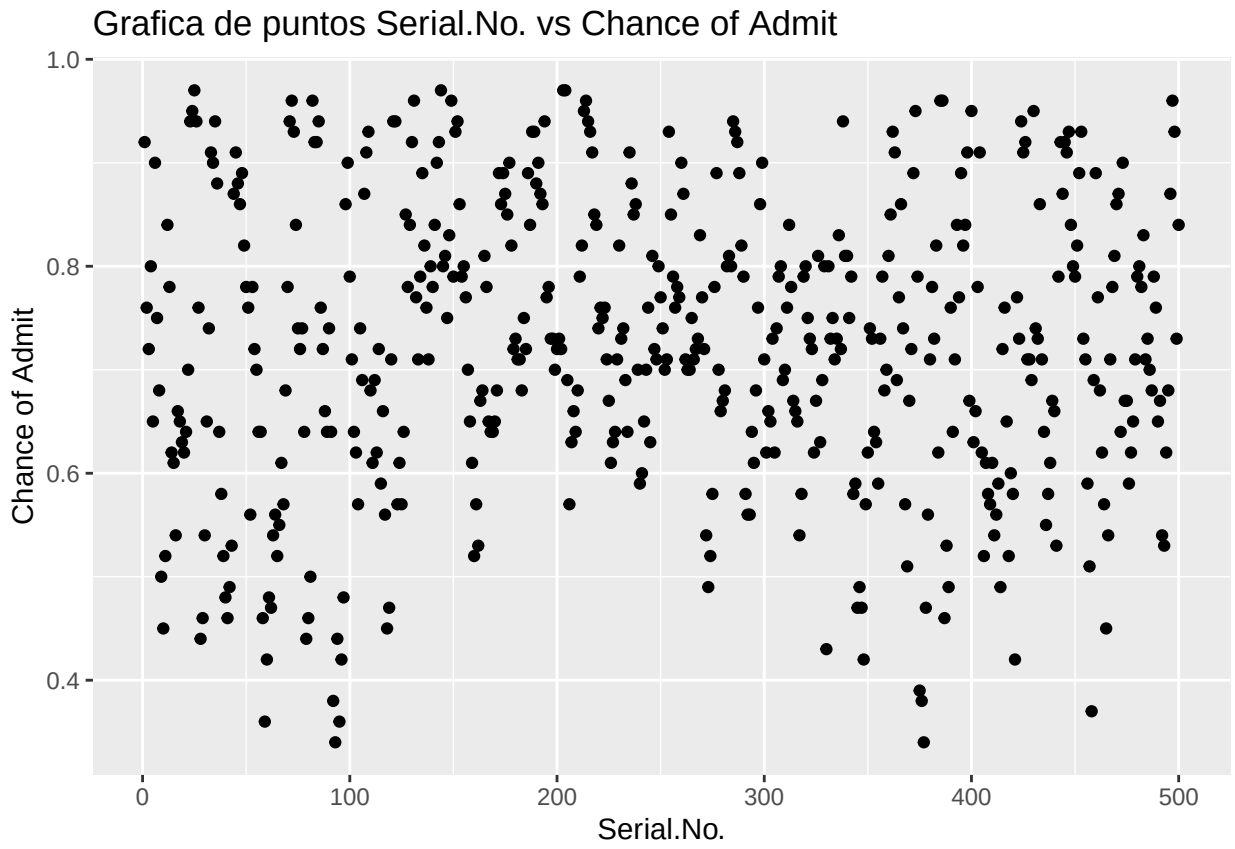
- La correlación más fuerte se encuentra entre `CGPA` y `Chance.of.Admit`, con un valor de 0.882. Esto indica una relación muy fuerte y positiva entre el rendimiento académico (CGPA) y la probabilidad de admisión.
- Las variables `TOEFL.Score` y `GRE.Score` también tienen una correlación considerablemente alta con `Chance.of.Admit`, con valores de 0.827 y 0.810, respectivamente. Esto indica que los puntajes altos en estos exámenes están asociados con una mayor probabilidad de admisión.

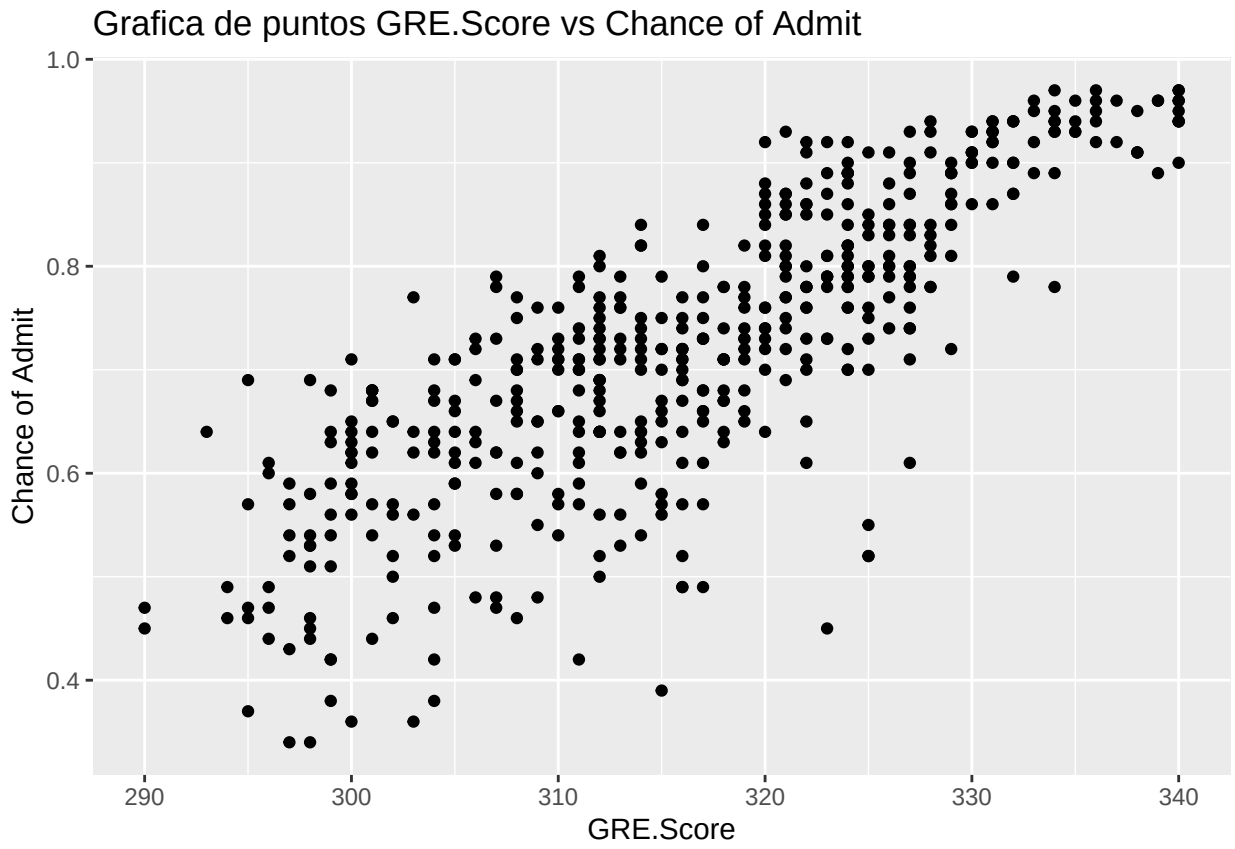
En general, los análisis estadísticos y la gráfica de correlación sugieren que las variables numéricas en el dataset están relacionadas y pueden ser indicadores importantes para predecir la probabilidad de admisión de los estudiantes.

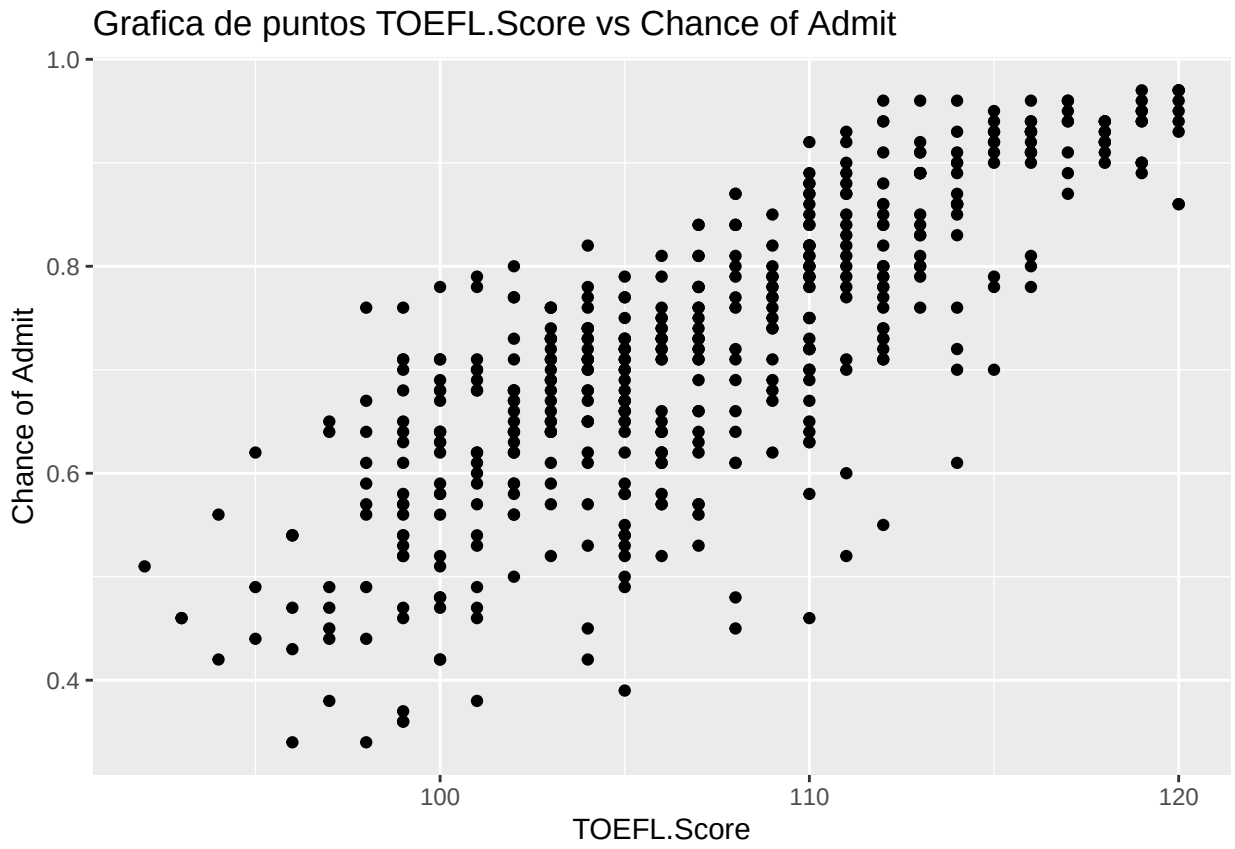
5. Realice un scatter plot (nube de puntos) de todas las variables numéricas contra la variable Chance of Admit.

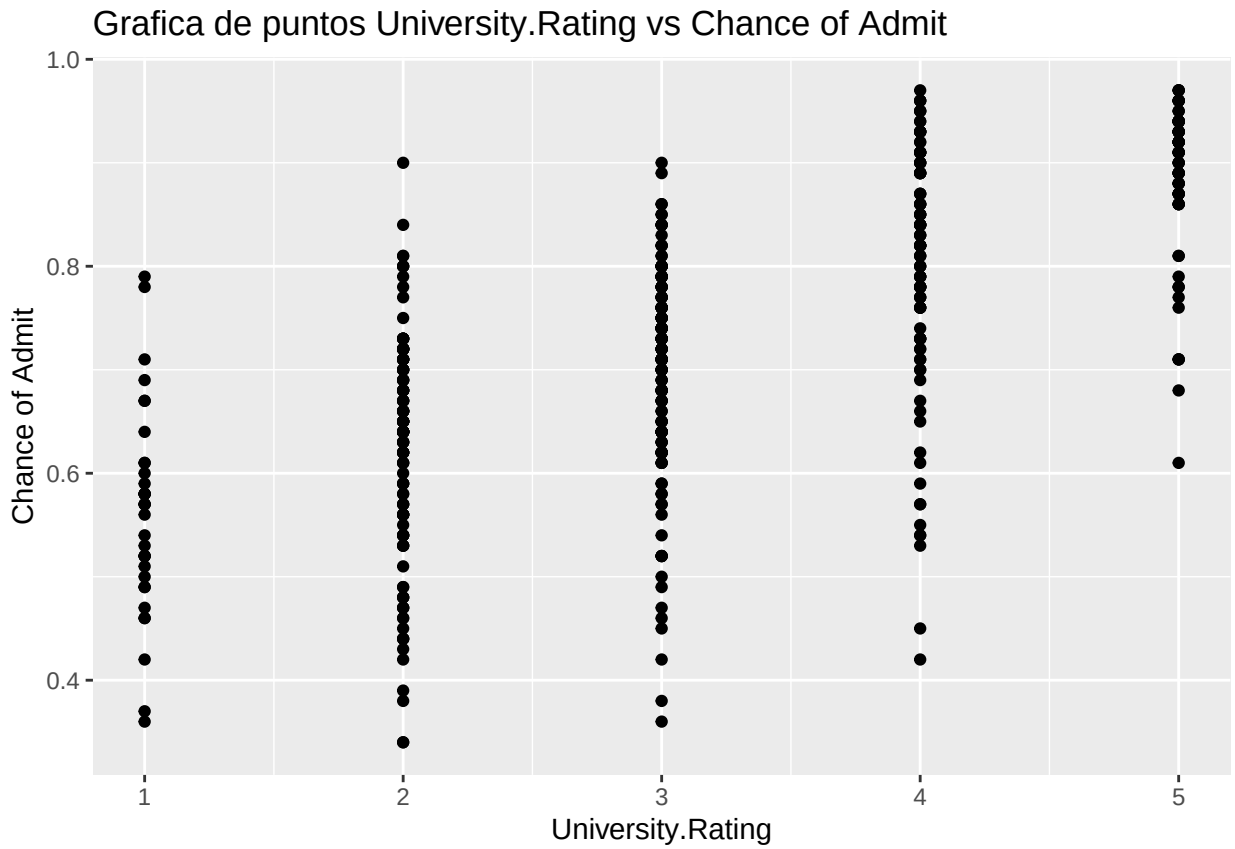
```
# Utilizar un bucle para crear gráficos para todas las variables
for (var in as.vector(colnames(dataset))){
  p <- ggplot(dataset, aes_string(x = var, y = "Chance.of.Admit")) +
    geom_point() +
    labs(title = paste("Grafica de puntos", var, "vs Chance of Admit"),
         x = var,
         y = "Chance of Admit")
  print(p)
}
```

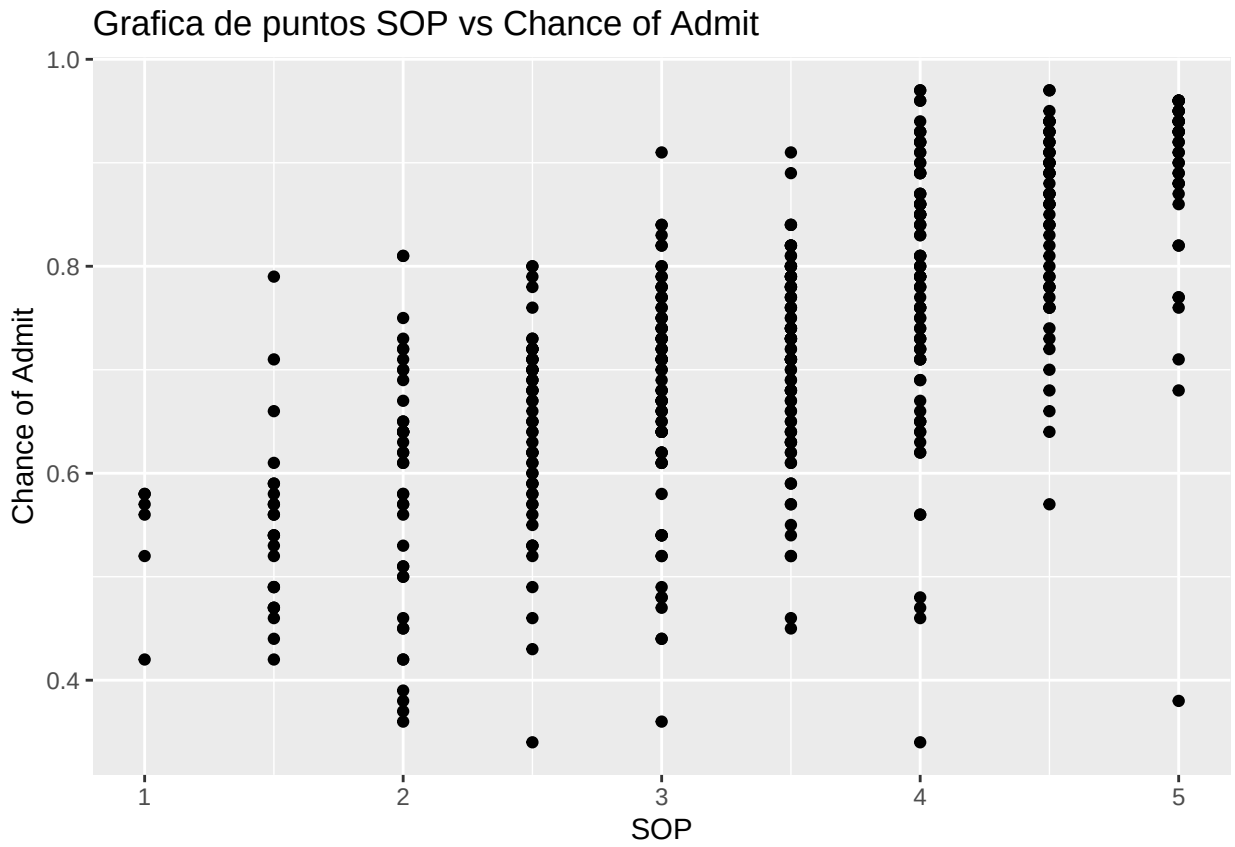
```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

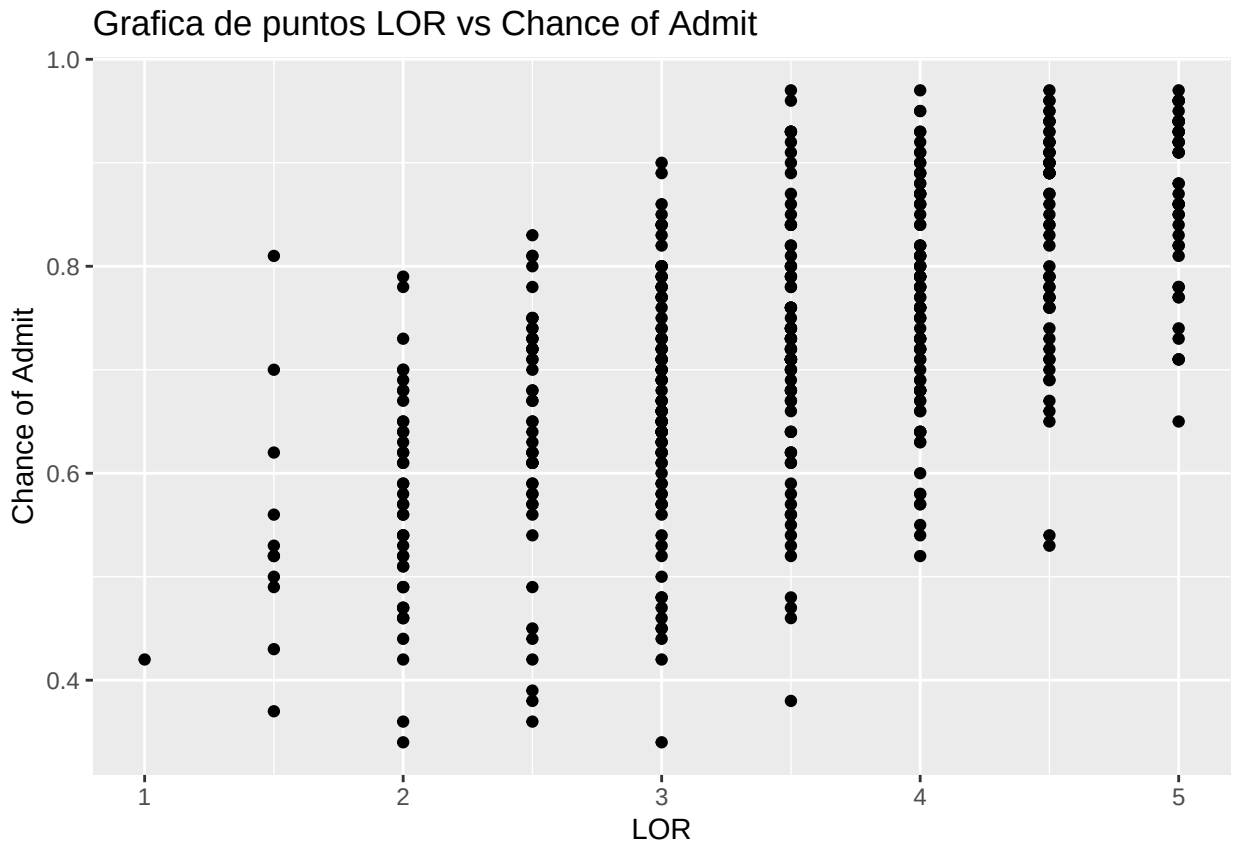

Grafica de puntos Serial.No. vs Chance of Admit
Chance of Admit
1.0
0.8
0.6
0.4
0
100
200
300
400
500
Serial.No.

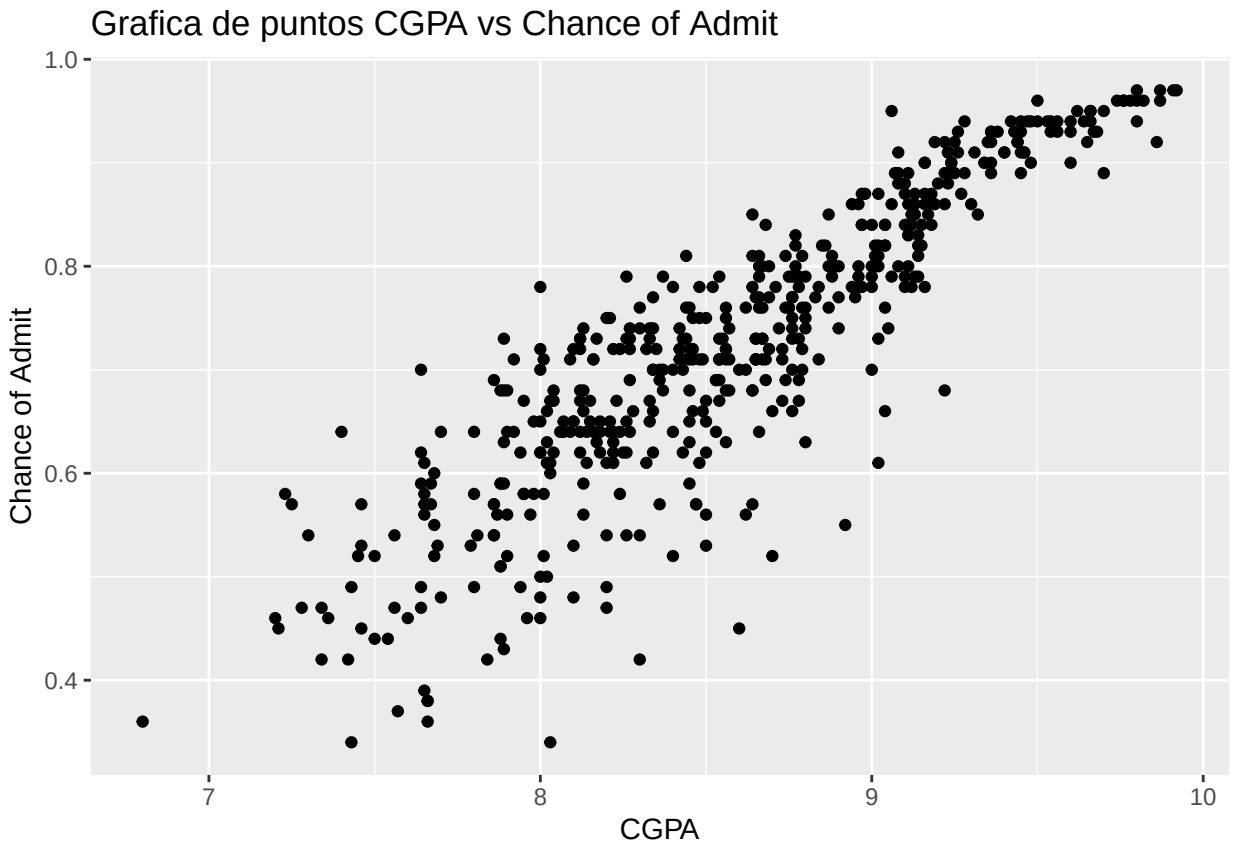

Grafica de puntos GRE.Score vs Chance of Admit
Chance of Admit
1.0
0.8
0.6
0.4
290
300
310
320
330
340
GRE.Score

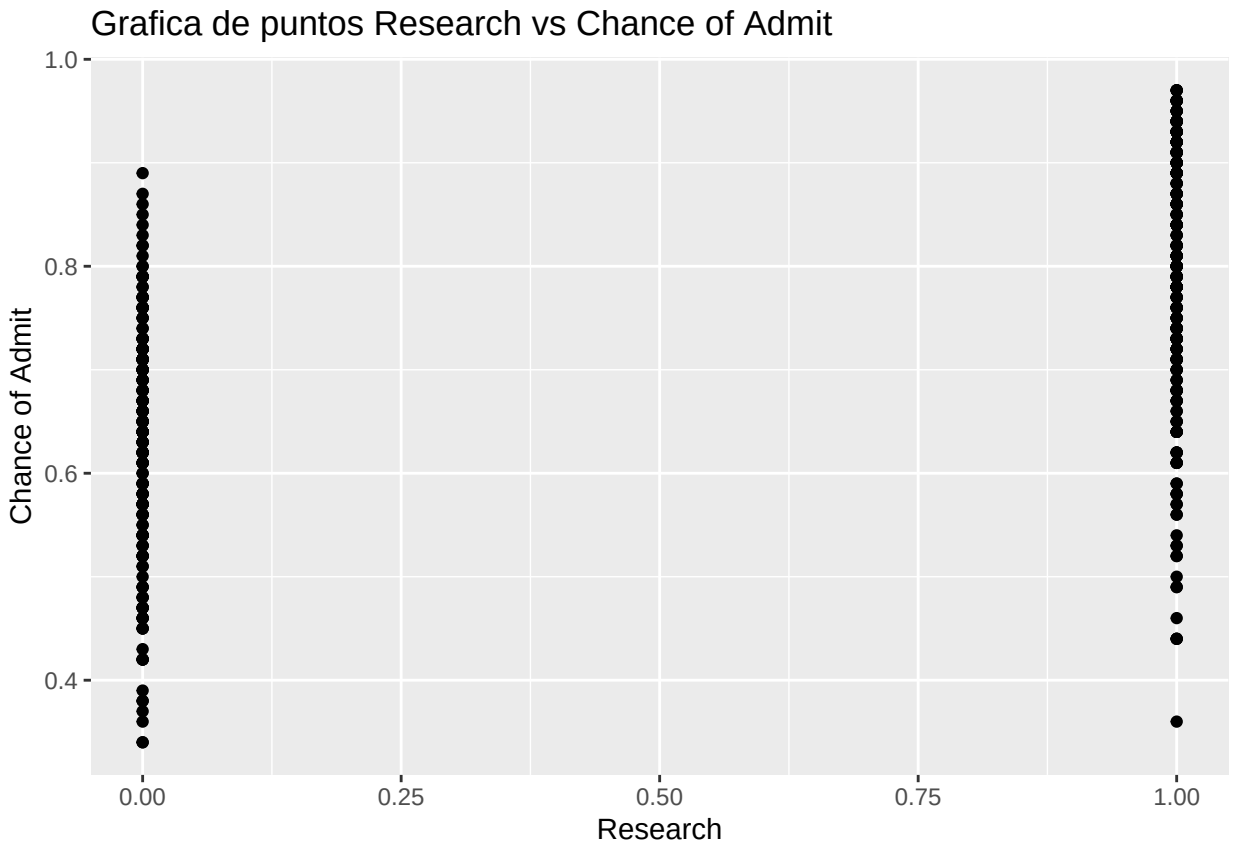

Grafica de puntos TOEFL.Score vs Chance of Admit
Chance of Admit
1.0
0.8
0.6
0.4
100
110
120
TOEFL.Score

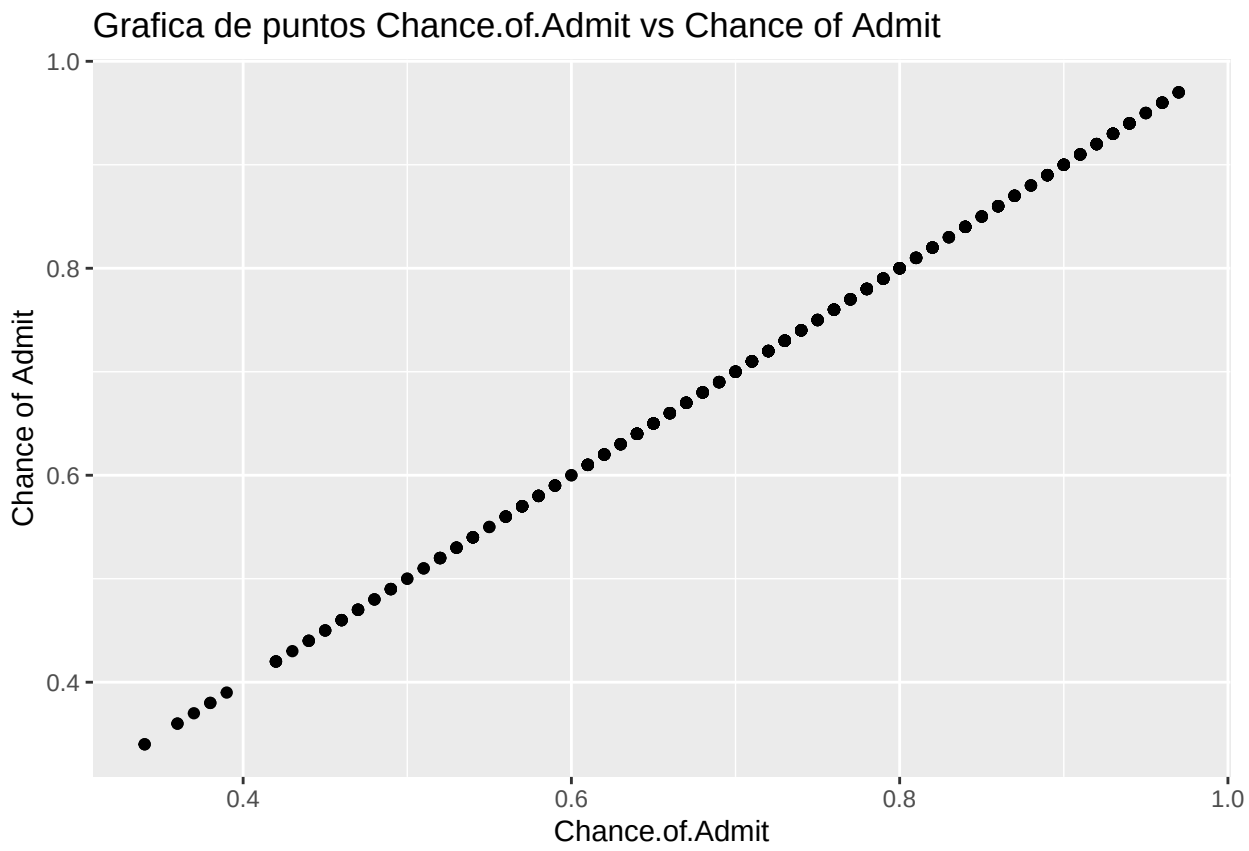

Grafica de puntos University.Rating vs Chance of Admit
Chance of Admit
1.0
0.8
0.6
0.4
1
2
3
4
5
University.Rating


Grafica de puntos SOP vs Chance of Admit
Chance of Admit
1.0
0.8
0.6
0.4
1
2
3
4
5
SOP


Grafica de puntos LOR vs Chance of Admit
Chance of Admit
1.0
0.8
0.6
0.4
1
2
3
4
5
LOR


Grafica de puntos CGPA vs Chance of Admit
Chance of Admit
1.0
0.8
0.6
0.4
7
8
9
10
CGPA


Grafica de puntos Research vs Chance of Admit
1.0
0.8
0.6
0.4
Chance of Admit
0.00
0.25
0.50
0.75
1.00
Research



6. Utilizando la función train y trainControl para crear un crossvalidation y le permita evaluar los siguientes modelos: • Chance of Admit ~ TOEFEL.Score. • Chance of Admit ~ CGPA. • Chance of Admit ~ GRE.Score. • Chance of Admit ~ TOEFEL.Score + CGPA. • Chance of Admit ~ TOEFEL.Score + GRE.Score. • Chance of Admit ~ GRE.Score + CGPA. • Chance of Admit ~ TOEFEL.Score + CGPA + GRE.Score. Posteriormente cree una lista ordenando de mejor a peor cual es el mejor modelo en predicción, recuerde que es necesario caclular el RMSE para poder armar correctamente la lista.

```
# Cargar la biblioteca caret
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.2.3
```

```
## Loading required package: lattice
```

```
# Establecer el método de control de entrenamiento
ctrl <- trainControl(method = "cv", number = 10)

# Especificar las fórmulas para los diferentes modelos
formulas <- list(
  "Chance.of.Admit ~ TOEFL.Score",
  "Chance.of.Admit ~ CGPA",
  "Chance.of.Admit ~ GRE.Score",
  "Chance.of.Admit ~ TOEFL.Score + CGPA",
  "Chance.of.Admit ~ TOEFL.Score + GRE.Score",
```

```
  "Chance.of.Admit ~ GRE.Score + CGPA",
  "Chance.of.Admit ~ TOEFL.Score + CGPA + GRE.Score"
)

# Inicializar una lista para almacenar los resultados de los modelos
results <- list()

# Utilizar un bucle para entrenar cada modelo y calcular el RMSE
for (i in 1:length(formulas)) {
  set.seed(123)
  model <- train(as.formula(formulas[[i]]), data = dataset, method = "lm", trControl = ctrl)
  results[[i]] <- list(
    Formula = formulas[[i]],
    RMSE = model$results$RMSE
  )
}

# Ordenar los resultados por RMSE en orden ascendente (menor a mayor)
#results <- do.call(rbind, results)
#results <- results[order(results$RMSE), ]

#results

# Crear un data frame con los resultados
results_df <- do.call(rbind, results)

results_df
```

```
##      Formula                                          RMSE
## [1,] "Chance.of.Admit ~ TOEFL.Score"                  0.08579917
## [2,] "Chance.of.Admit ~ CGPA"                         0.06643456
## [3,] "Chance.of.Admit ~ GRE.Score"                    0.08253551
## [4,] "Chance.of.Admit ~ TOEFL.Score + CGPA"           0.06376555
## [5,] "Chance.of.Admit ~ TOEFL.Score + GRE.Score"      0.07675989
## [6,] "Chance.of.Admit ~ GRE.Score + CGPA"             0.06311976
## [7,] "Chance.of.Admit ~ TOEFL.Score + CGPA + GRE.Score" 0.06241351
```

```
typeof(results_df)
```

```
## [1] "list"
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Convertir la lista en un data frame
results_df <- bind_rows(results)

# Imprimir el data frame results_df
print(results_df)
```

```
## # A tibble: 7 x 2
##   Formula                                          RMSE
##   <chr>                                           <dbl>
## 1 Chance.of.Admit ~ TOEFL.Score                  0.0858
## 2 Chance.of.Admit ~ CGPA                         0.0664
## 3 Chance.of.Admit ~ GRE.Score                    0.0825
## 4 Chance.of.Admit ~ TOEFL.Score + CGPA           0.0638
## 5 Chance.of.Admit ~ TOEFL.Score + GRE.Score      0.0768
## 6 Chance.of.Admit ~ GRE.Score + CGPA             0.0631
## 7 Chance.of.Admit ~ TOEFL.Score + CGPA + GRE.Score 0.0624
```

```
# Seleccionar la fila con el valor mínimo de RMSE
best_model <- filter(results_df, RMSE == min(RMSE))

# Imprimir el mejor modelo
print(best_model)
```

```
## # A tibble: 1 x 2
##   Formula                                          RMSE
##   <chr>                                           <dbl>
## 1 Chance.of.Admit ~ TOEFL.Score + CGPA + GRE.Score 0.0624
```