Activity 5 - Mini-competition Explorations

```
library("knitr")
library("kableExtra")
library("tidyverse")
library("tidymodels")
library("GGally")
library("psych")
library("ggfortify")
```

Import Data

```
allendale_students <- readr::read_csv("./data/allendale-students.csv")

knitr::kable(head(allendale_students))
```

distance	scholarship	parents	car	housing	major	debt
40	1532	0.440	6	off campus	STEM	26389
30	7479	0.265	7	on campus	STEM	21268
130	2664	0.115	3	on campus	business	32312
120	1998	0.325	9	on campus	business	28539
30	1462	0.105	10	off campus	other	34867
0	3053	0.335	9	off campus	STEM	18193

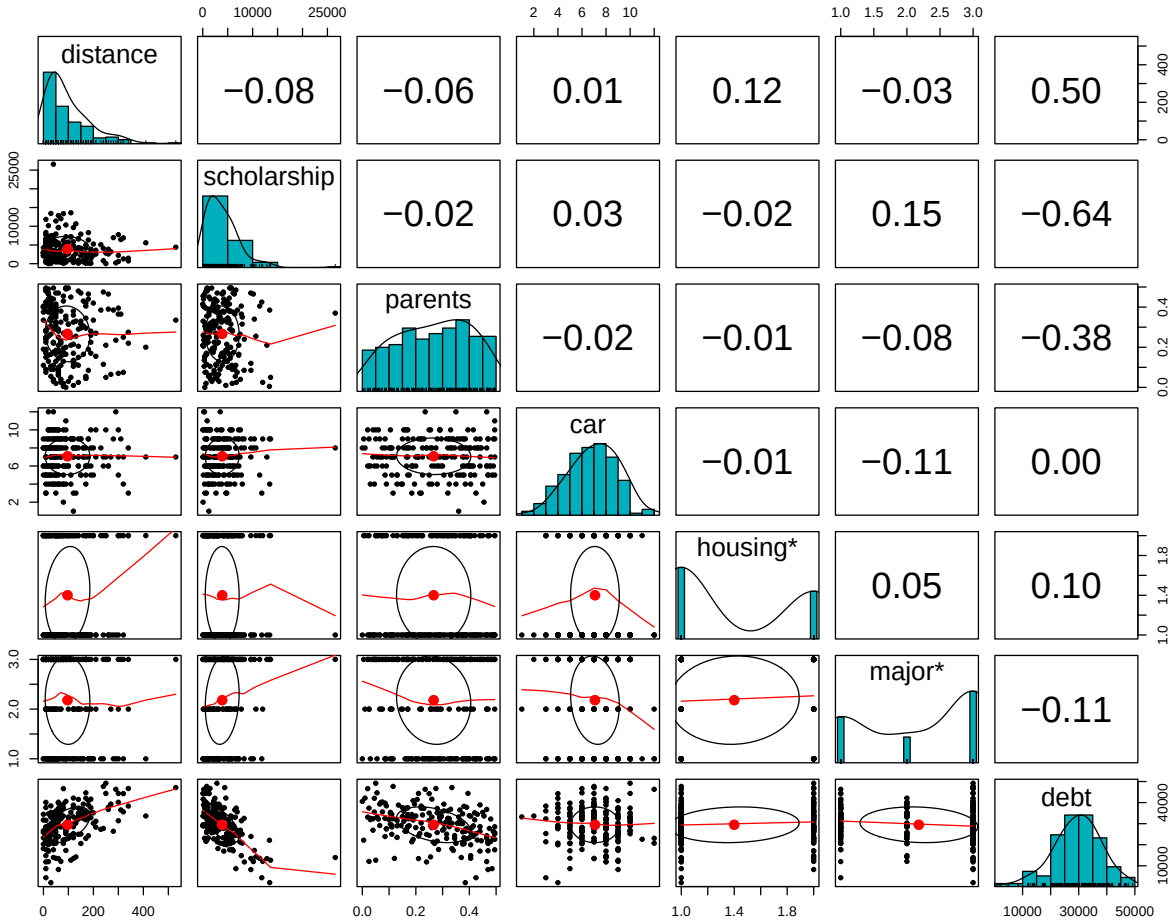
Initial Analysis

```
kable_styling(knitr::kable(summary(allendale_students), booktabs = T, format="latex"), position="center
```

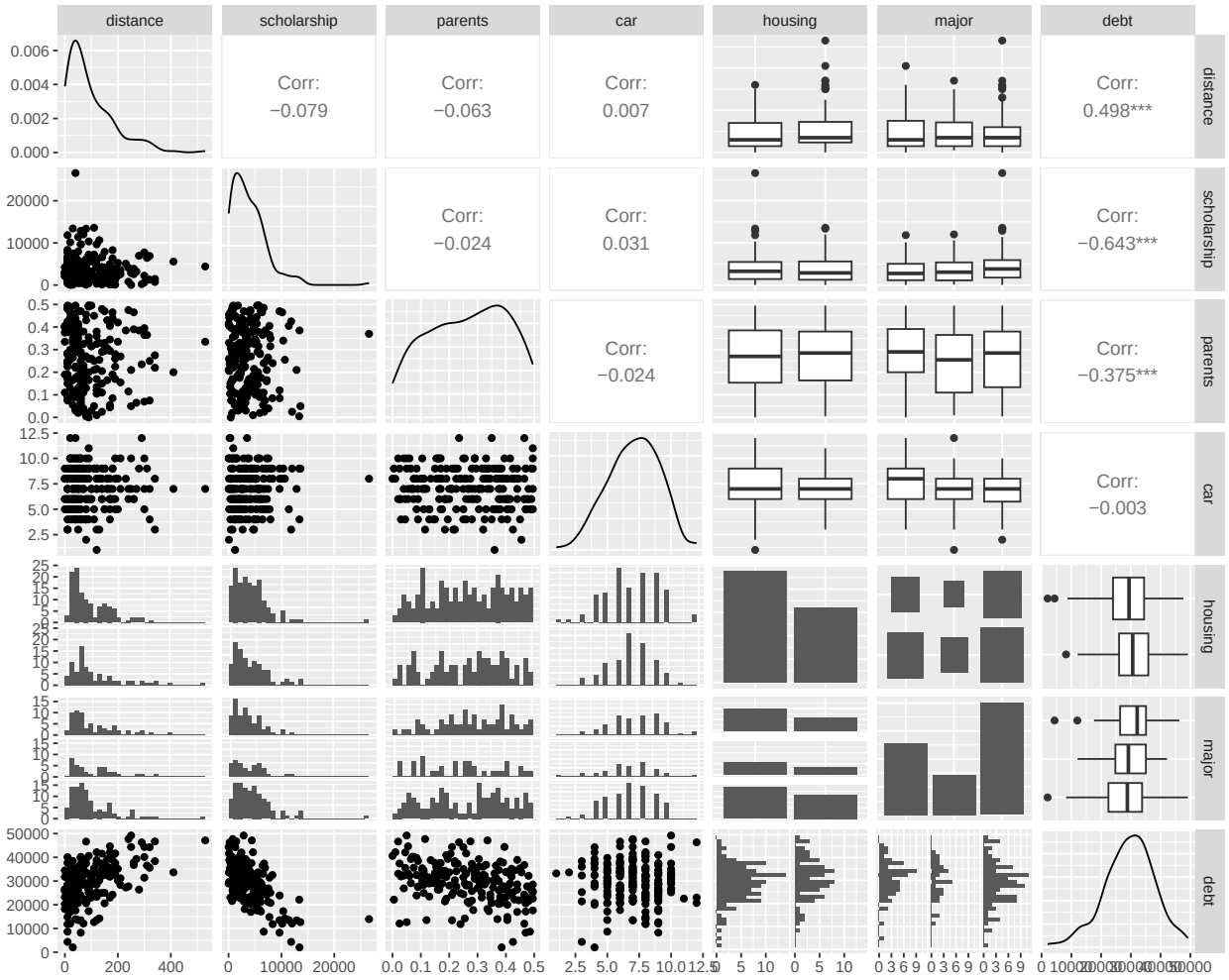
```
psych::pairs.panels(
  allendale_students,
  hist.col = "#00AFBB",
  method= "pearson",
```

distance	scholarship	parents	car	housing	major	debt
Min. : 0.00	Min. : 25	Min. :0.0000	Min. : 1.00	Length:200	Length:200	Min. : 2019
1st Qu.: 30.00	1st Qu.: 1312	1st Qu.:0.1588	1st Qu.: 6.00	Class :character	Class :character	1st Qu.:2423
Median : 70.00	Median : 3202	Median :0.2800	Median : 7.00	Mode :character	Mode :character	Median :298
Mean : 96.55	Mean : 3899	Mean :0.2666	Mean : 7.08	NA	NA	Mean :29473
3rd Qu.:140.00	3rd Qu.: 5504	3rd Qu.:0.3812	3rd Qu.: 9.00	NA	NA	3rd Qu.:3502
Max. :530.00	Max. :26574	Max. :0.4950	Max. :12.00	NA	NA	Max. :49196

```
  density = TRUE,
  ellipses = TRUE
)
```




```
allendale_students %>%
  ggpairs()
```



From the above visualization, we can observe that the observations in distance and scholarship variables are skewed towards right. Also, we can observe that the variable debt has some correlation with distance, scholarship and parents variables.

Perform Single Linear Regression

Lets use the `lm` function to fit the linear model where `y` is debt and `x` is distance, scholarship, parents, car, and housing

SLR: debt and distance

```
m_distance <- lm(debt ~ distance, data = allendale_students)
tidy(m_distance)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  24911.     768.       32.4  4.07e-81
## 2 distance        47.3      5.85      8.08 6.16e-14
```

```
glance(m_distance)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squa~1 sigma stati~2  p.value    df logLik   AIC   BIC devia~3
##       <dbl>        <dbl> <dbl>   <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1     0.248        0.244 7376.    65.3 6.16e-14     1 -2064. 4134. 4144. 1.08e10
## # ... with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
## #   variable names 1: adj.r.squared, 2: statistic, 3: deviance
```

SLR: debt and scholarship

```
m_scholarship <- lm(debt ~ scholarship, data = allendale_students)
tidy(m_scholarship)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept) 35736.     702.         50.9 1.19e-115
## 2 scholarship    -1.61     0.136     -11.8 1.02e- 24
```

```
glance(m_scholarship)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squa~1 sigma stati~2  p.value    df logLik   AIC   BIC devia~3
##       <dbl>        <dbl> <dbl>   <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1     0.413        0.410 6515.    140. 1.02e-24     1 -2039. 4084. 4094.  8.40e9
## # ... with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
## #   variable names 1: adj.r.squared, 2: statistic, 3: deviance
```

SLR: debt and parents

```
m_parents <- lm(debt ~ parents, data = allendale_students)
tidy(m_parents)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   35587.     1209.      29.4  3.03e-74
## 2 parents      -22932.     4023.      -5.70 4.30e- 8
```

```
glance(m_parents)
```

```
## # A tibble: 1 x 12
##   r.squ~1 adj.r~2 sigma stati~3 p.value    df logLik   AIC   BIC devia~4 df.re~5
##     <dbl>   <dbl> <dbl>   <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>   <int>
## 1   0.141   0.137 7884.    32.5 4.30e-8     1 -2077. 4161. 4171. 1.23e10     198
## # ... with 1 more variable: nobs <int>, and abbreviated variable names
## #   1: r.squared, 2: adj.r.squared, 3: statistic, 4: deviance, 5: df.residual
```

SLR: debt and car

```
m_car <- lm(debt ~ car, data = allendale_students)
tidy(m_car)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  29554.      2192.   13.5    7.91e-30
## 2 car            -11.4      298.   -0.0384 9.69e- 1
```

```
glance(m_car)
```

```
## # A tibble: 1 x 12
##    r.squared adj.r.squa~1 sigma stati~2 p.value    df logLik   AIC   BIC devia~3
##        <dbl>        <dbl> <dbl>   <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1 0.00000746     -0.00504 8506. 0.00148   0.969     1 -2092. 4191. 4201. 1.43e10
## # ... with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
## #   variable names 1: adj.r.squared, 2: statistic, 3: deviance
```

Multiple Linear Regression

Fit the multiple linear regression with debt as dependent variables and distance, scholarship and parents as independent variables.

```
m_mlr_dsp <- lm(debt ~ distance + scholarship + parents , data = allendale_students)
tidy(m_mlr_dsp)
```

```
## # A tibble: 4 x 5
##   term         estimate std.error statistic   p.value
##   <chr>           <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)  37510.    852.          44.0 1.47e-103
## 2 distance        40.5     3.43        11.8 1.20e- 24
## 3 scholarship     -1.54    0.0901     -17.1 8.04e- 41
## 4 parents     -22216.   2201.         -10.1 1.50e- 19
```

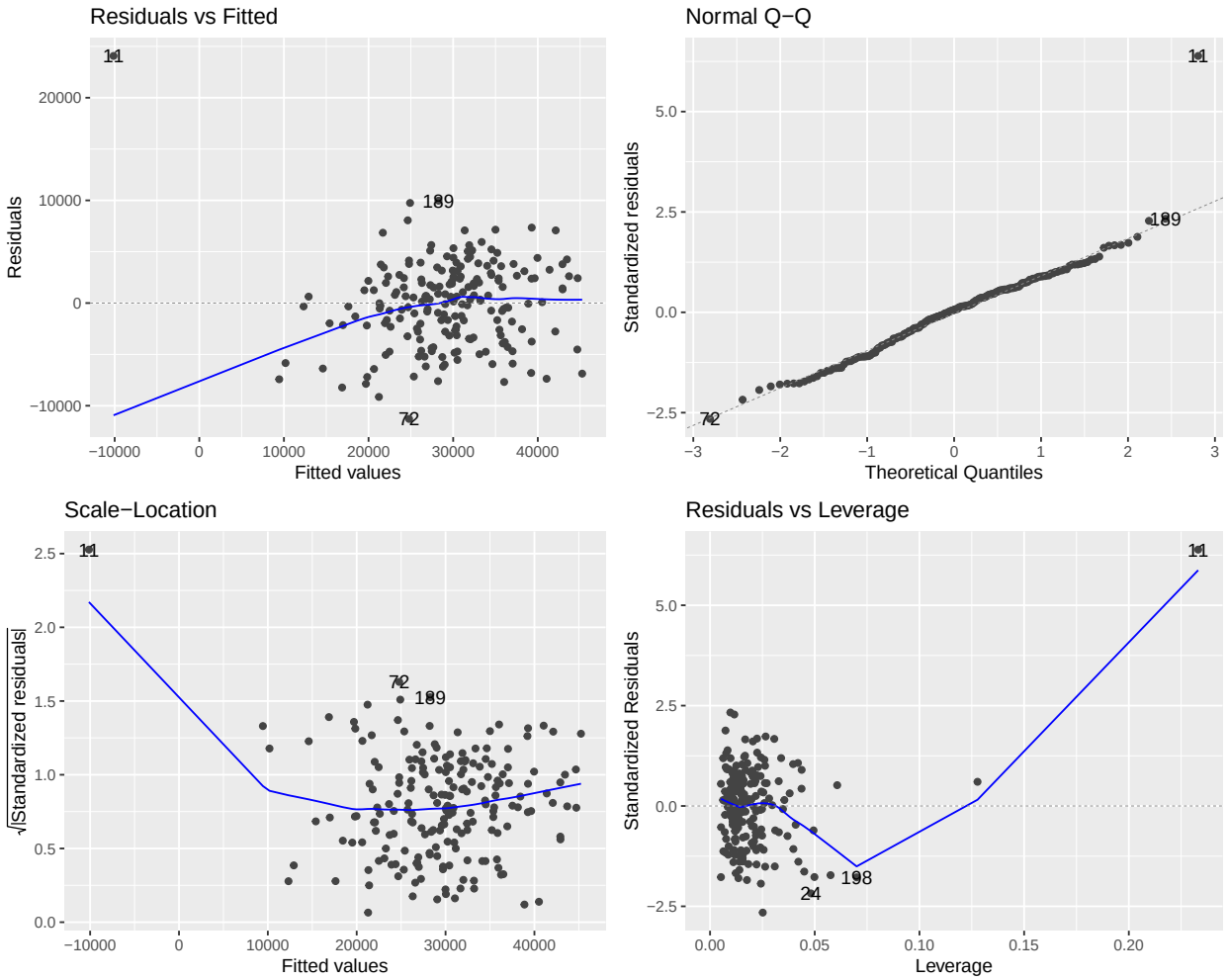
Access model fit using `glance`:

```
glance(m_mlr_dsp)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squa~1 sigma stati~2  p.value    df logLik   AIC   BIC devia~3
##       <dbl>        <dbl> <dbl>   <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1     0.747        0.743 4304.    193. 3.62e-58     3 -1955. 3920. 3937.  3.63e9
## # ... with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
## #   variable names 1: adj.r.squared, 2: statistic, 3: deviance
```

Now, lets access the model diagnostics using the `ggplot2::autoplot` function.

```
ggplot2::autoplot(m_mlr_dsp)
```



In the above diagnostics plot, we can observe that observation 11 is clearly an outlier. Lets use `broom::augment` to further analyze the outlier.

```
# https://broom.tidymodels.org/reference/augment.lm.html
augment_allendale_students <- broom::augment(m_mlr_dsp, data=allendale_students)
```

Looking into `augment_allendale_students` variable, it looks like observation `11` is clearly an outlier. lets remove it from the data:

```
data <- allendale_students %>%
  filter(!row_number() %in% c(11))
```

Now, lets fit the multiple linear regression again:

```
m_mlr_dsp <- lm(debt ~ distance + scholarship + parents , data = data)
tidy(m_mlr_dsp)
```

```
## # A tibble: 4 x 5
```

```
##   term         estimate std.error statistic   p.value
##   <chr>           <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)  38830.    782.          49.6 1.41e-112
## 2 distance        40.6     3.06        13.3 4.93e- 29
## 3 scholarship     -1.86    0.0914     -20.3 5.65e- 50
## 4 parents     -23241.   1969.         -11.8 1.32e- 24
```

```
glance(m_mlr_dsp)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squa~1 sigma stati~2  p.value    df logLik   AIC   BIC devia~3
##       <dbl>        <dbl> <dbl>   <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1     0.796        0.793 3840.    253. 5.16e-67     3 -1923. 3855. 3872.  2.87e9
## # ... with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
## #   variable names 1: adj.r.squared, 2: statistic, 3: deviance
```

We can observed imporved model after removing the outlier. Now lets try to fit the MLR with different interation:

```
m_mlr_1 <- lm(debt ~ distance * car + scholarship + parents, data = data)
glance(m_mlr_1)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squa~1 sigma stati~2  p.value    df logLik   AIC   BIC devia~3
##       <dbl>        <dbl> <dbl>   <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1     0.798        0.793 3840.    152. 4.87e-65     5 -1922. 3857. 3880.  2.85e9
## # ... with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
## #   variable names 1: adj.r.squared, 2: statistic, 3: deviance
```

```
m_mlr_2 <- lm(debt ~ distance * scholarship * parents * housing, data = data)
glance(m_mlr_2)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squa~1 sigma stati~2  p.value    df logLik   AIC   BIC devia~3
##       <dbl>        <dbl> <dbl>   <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1     0.808        0.792 3848.    51.2 3.20e-57    15 -1917. 3868. 3924.  2.71e9
## # ... with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
## #   variable names 1: adj.r.squared, 2: statistic, 3: deviance
```

```
m_mlr_3 <- lm(debt ~ distance * scholarship * parents * major, data = data)
glance(m_mlr_3)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squa~1 sigma stati~2  p.value    df logLik   AIC   BIC devia~3
##       <dbl>        <dbl> <dbl>   <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1     0.835        0.813 3643.    38.5 1.96e-56    23 -1901. 3853. 3935.  2.32e9
## # ... with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
## #   variable names 1: adj.r.squared, 2: statistic, 3: deviance
```

The model `m_mlr_3` seems to be the best fit.

Next, let investigate if performing normalization can boost our model:

```r
norm_data <- data %>%
  mutate(log_distance = log(distance)) %>%
  mutate(log_scholarship = log(scholarship))

# Remove inf values
norm_data$log_distance[!is.finite(norm_data$log_distance)] <- 0

# fit model
m_mlr_4 <- lm(debt ~ log_distance + scholarship + parents, data = norm_data)
glance(m_mlr_4)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squa~1 sigma stati~2  p.value    df logLik   AIC   BIC devia~3
##       <dbl>        <dbl> <dbl>   <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1     0.741        0.737 4325.    186. 5.90e-57     3 -1946. 3903. 3919.  3.65e9
## # ... with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
## #   variable names 1: adj.r.squared, 2: statistic, 3: deviance
```

No improvements, The model `m_mlr_3` produces better result.