HW5-Spandan-Maaheshwari

Spandan Maaheshwari

2022-11-20

Problem 1 ->

Name of the student whose miniposter I chose: Arya Dhorajiya

The original source of the dataset used in the miniposter: https://www.kaggle.com/datasets/dorinaferencsik/outdoor-cycling-metrics?select=cycling_metrics_clean.csv

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2

## Warning: package 'ggplot2' was built under R version 4.2.2

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(ggplot2)

bike_riders <- read.csv("/Users/SPANDAN/DS 5110/cycling_metrics_clean.csv")
bike_riders <- bike_riders |> dplyr::filter(age_group==1) |> mutate(avg_speed_km = 3.6 * average_speed,

head(bike_riders,10)
```

```
##                               hashed_id age_group average_speed distance
## 1  fa498f22-edef-4a9d-af00-ef07e2585072         1         7.099 102.9770
## 2  fa498f22-edef-4a9d-af00-ef07e2585072         1         7.040  64.0465
## 3  fa498f22-edef-4a9d-af00-ef07e2585072         1         7.371  38.8537
## 4  fa498f22-edef-4a9d-af00-ef07e2585072         1         8.119  39.0758
## 5  fa498f22-edef-4a9d-af00-ef07e2585072         1         6.839  51.2323
## 6  fa498f22-edef-4a9d-af00-ef07e2585072         1         6.513 201.6820
## 7  fa498f22-edef-4a9d-af00-ef07e2585072         1         7.311 100.3960
## 8  fa498f22-edef-4a9d-af00-ef07e2585072         1         8.004  33.4741
## 9  fa498f22-edef-4a9d-af00-ef07e2585072         1         6.890  38.7259
## 10 fa498f22-edef-4a9d-af00-ef07e2585072         1         6.765  72.9834
##    elapsed_time highest_elevation lowest_elevation max_speed moving_time
```

```
## 1         15212            476.6            241.6     77.76       14505
## 2         10964            394.8            221.2     61.56        9097
## 3          5271             68.4            -27.4     45.72        5271
## 4          6634             57.8            -28.2     60.12        4813
## 5          7491            228.6            -10.2     62.28        7491
## 6         35048            501.6             71.6     64.80       30966
## 7         15037            297.2             60.8     70.92       13732
## 8          4989            246.0             55.8     49.32        4182
## 9          5637            260.6             54.2     55.44        5621
## 10        11396            408.2            157.6     64.08       10789
##        start_date_local total_elevation_gain avg_speed_km
## 1  2019-03-04 09:07:04                  968      25.5564
## 2  2019-02-28 09:37:24                  768      25.3440
## 3  2019-02-25 12:19:24                  162      26.5356
## 4  2019-02-25 10:20:43                   79      29.2284
## 5  2019-02-24 14:35:25                  608      24.6204
## 6  2019-02-20 08:00:40                 1941      23.4468
## 7  2019-02-17 09:42:05                  967      26.3196
## 8  2019-02-16 11:36:57                  255      28.8144
## 9  2019-02-14 12:27:56                  446      24.8040
## 10 2019-02-10 09:45:08                  780      24.3540
```

Explanation for Problem 1 ->

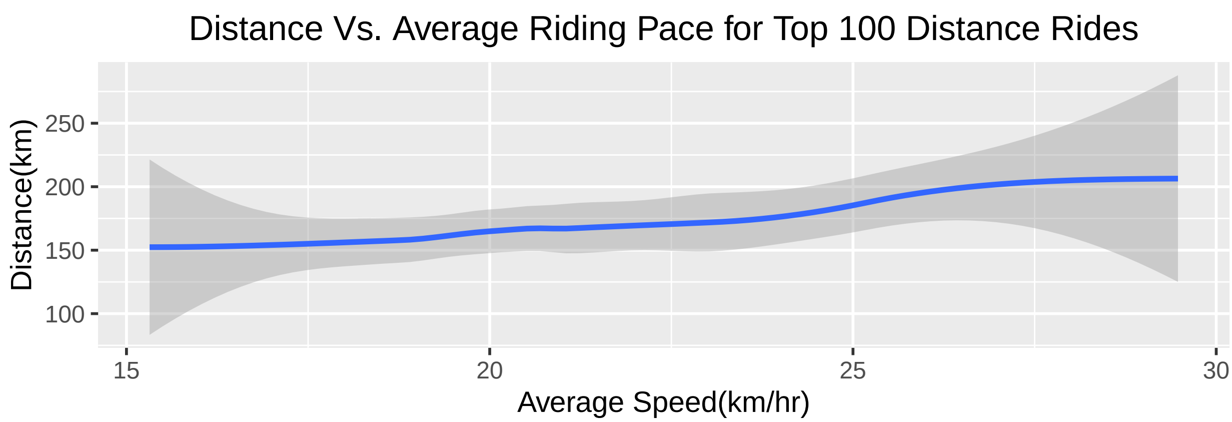
Avg. Pace, Distance Traveled, Total Elevation Gained, Time Duration, and Age Group are some of the common metrics in cycling terms that were recorded in the dataset of cyclists.

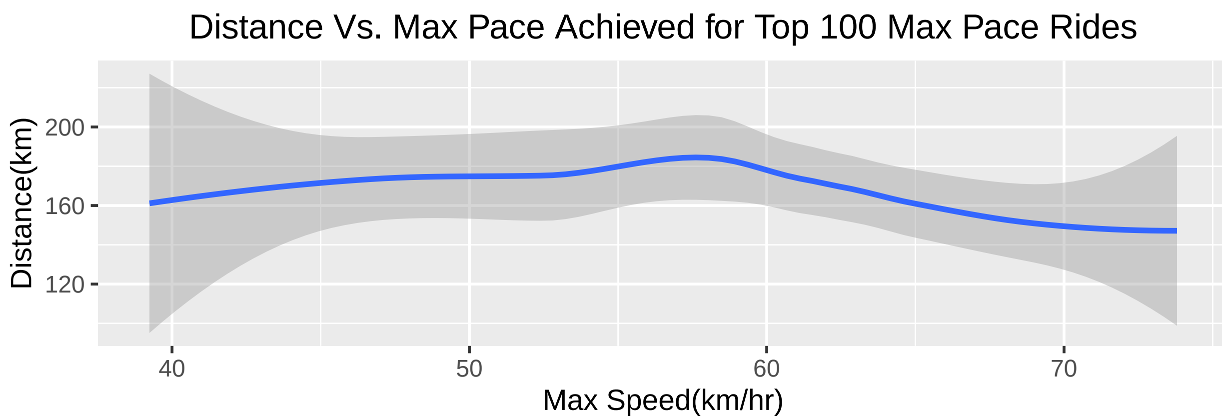
In order to get a good approximation of high performance, I had to limit my age group for visualization since 18-35 yrs age group made up roughly 85% of the overall data . Additionally, converted speed and distance into Km/hr and Kms, respectively, as they were measured in m/s. The data wasn't further cleaned nor arranged as it was already into a tidy format.

Problem 2 ->

```
bike_riders <- bike_riders |> arrange(desc(distance))
bike_riders <- head(bike_riders,100)

g1 <- bike_riders %>%
  ggplot(aes(x=avg_speed_km,y=distance)) +
    geom_smooth() + labs(x="Average Speed(km/hr)", y = "Distance(km)",
       title = "Distance Vs. Average Riding Pace for Top 100 Distance Rides") +
         theme(plot.title = element_text(hjust = 0.5))


g2 <- bike_riders%>%
   ggplot(aes(x=max_speed,y=distance)) +
    geom_smooth() + labs(x="Max Speed(km/hr)", y = "Distance(km)",
       title = "Distance Vs. Max Pace Achieved for Top 100 Max Pace Rides") +
         theme(plot.title = element_text(hjust = 0.5))



gridExtra::grid.arrange(g1, g2, nrow=2)
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```





Explanation for Problem 2:

From the first plot it is evident that The cyclist's average speed is between 15 and 30 km/hr, and as their average speed increases, so does the distance covered throughout the trip.

The cyclist's maximum speed is recorded between 40 and 77 km/h, and the trend curve shows a peak at the point where he covers the greatest distance, 183 km, at a maximum speed of 57 km/h. The fact that the distance typically decreases as the maximum speed rises indicates that the rider may be engaging in a high-pace, short-distance ride.

Problem 3 ->

```
library(tidyverse)
library(modelr)
```

```
## Warning: package 'modelr' was built under R version 4.2.2
```

```
library(mlbench)
```

```
## Warning: package 'mlbench' was built under R version 4.2.2
```

```
data(PimaIndiansDiabetes2)
pima <- as_tibble(PimaIndiansDiabetes2)
pima <- na.omit(pima)

pima$diabetes <- as.factor(pima$diabetes)
```

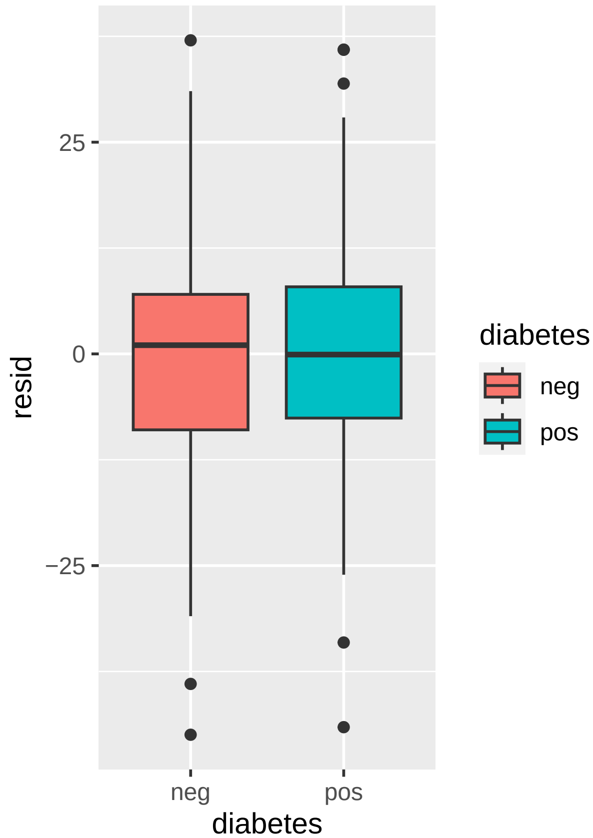
```r
fit <- lm(pressure ~ diabetes, data = pima)

summary(fit)
```

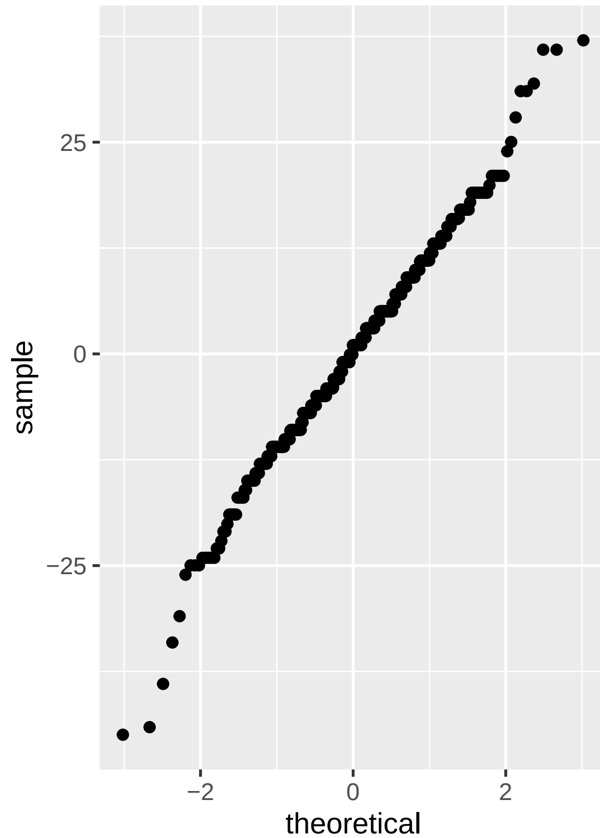
```
##
## Call:
## lm(formula = pressure ~ diabetes, data = pima)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.969  -8.077   1.031   7.923  37.031
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  68.9695     0.7585  90.927  < 2e-16 ***
## diabetespos   5.1075     1.3172   3.878 0.000124 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.28 on 390 degrees of freedom
## Multiple R-squared:  0.03712,    Adjusted R-squared:  0.03465
## F-statistic: 15.04 on 1 and 390 DF,  p-value: 0.0001237
```

Model diagnostics

A effective technique to evaluate a model visually and look for model assumption breaches is to plot the residuals (errors)

```r
g1 <- pima %>%
  add_residuals(fit, "resid") %>%
    ggplot(aes(x=diabetes,y=resid, fill=diabetes)) +
      geom_boxplot() + labs(x="diabetes")

g2 <- pima %>%
  add_residuals(fit, "resid") %>%
    ggplot(aes(sample=resid)) + geom_qq()

gridExtra::grid.arrange(g1, g2, ncol=2)
```





The residuals are roughly normal, as seen by the QQ-plot. But we find a significant outliers that needs to be eliminated:

```
outlier <- pima %>%
  add_residuals(fit, "resid") %>%
  filter(resid < -30)

outlier
```

```
## # A tibble: 5 x 10
##   pregnant glucose pressure triceps insulin  mass pedigree   age diabetes resid
##      <dbl>   <dbl>    <dbl>   <dbl>   <dbl> <dbl>    <dbl> <dbl> <fct>    <dbl>
## 1        0     137       40      35     168  43.1    2.29     33 pos      -34.1
## 2        1     103       30      38      83  43.3    0.183    33 neg      -39.0
## 3        1      88       30      42      99  55      0.496    26 pos      -44.1
## 4        1      89       24      19      25  27.8    0.559    21 neg      -45.0
## 5        1     109       38      18     120  23.1    0.407    26 neg      -31.0
```

```
pima1 <- anti_join(pima, outlier)
```

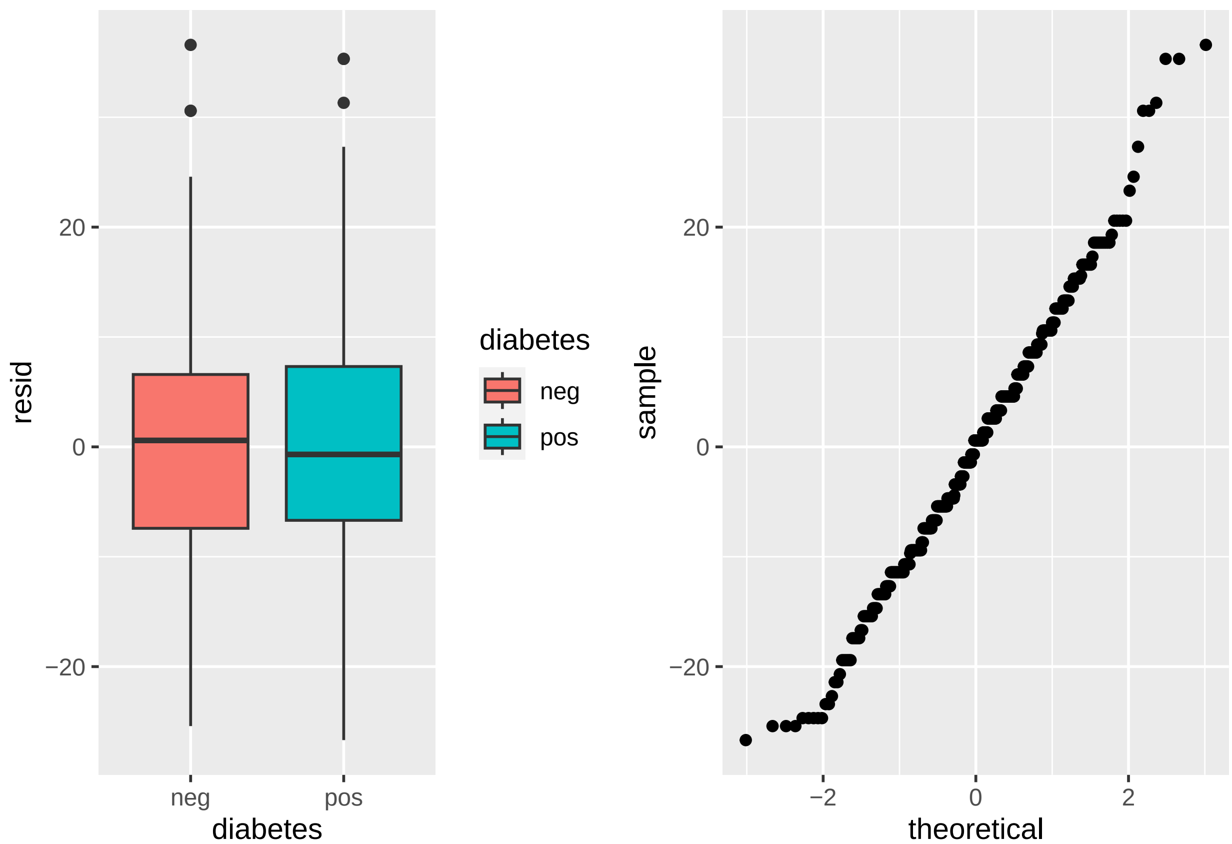
```
## Joining, by = c("pregnant", "glucose", "pressure", "triceps", "insulin",
## "mass", "pedigree", "age", "diabetes")
```

```
fit1 <- lm(pressure ~ diabetes, data = pima1)
summary(fit1)
```

```
##
## Call:
## lm(formula = pressure ~ diabetes, data = pima1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.688  -7.413   0.587   7.312  36.587
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  69.4131     0.7158  96.977  < 2e-16 ***
## diabetespos   5.2744     1.2446   4.238 2.83e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.52 on 385 degrees of freedom
## Multiple R-squared:  0.04457,    Adjusted R-squared:  0.04209
## F-statistic: 17.96 on 1 and 385 DF,  p-value: 2.827e-05
```

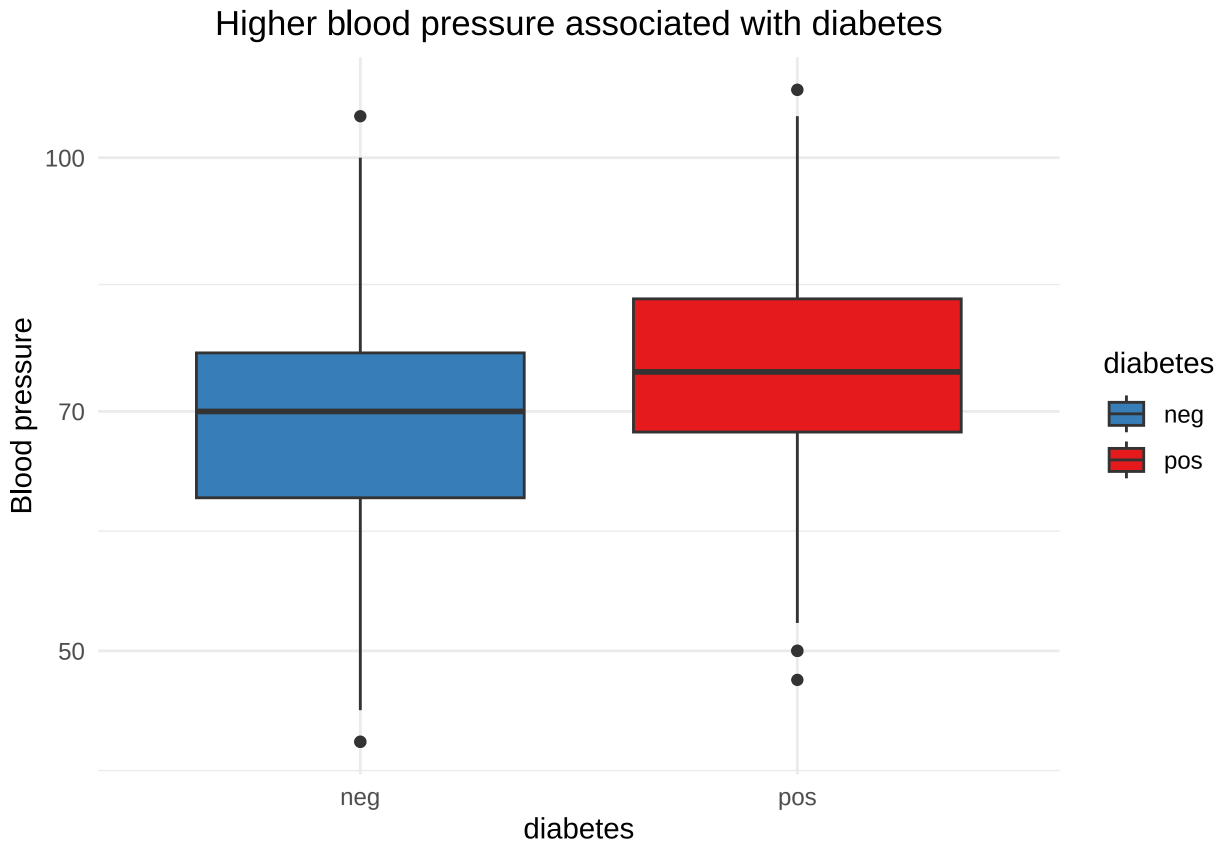
We plot residuals once more after deleting the outlier and re-fitting the model:

```r
g1 <- pima1 %>%
  add_residuals(fit1, "resid") %>%
    ggplot(aes(x=diabetes,y=resid, fill=diabetes)) +
      geom_boxplot() + labs(x="diabetes")

g2 <- pima1 %>%
  add_residuals(fit1, "resid") %>%
    ggplot(aes(sample=resid)) + geom_qq()

gridExtra::grid.arrange(g1, g2, ncol=2)
```



Like before, there are no violations of the model assumptions, the residuals appear to be approximately normal, according to the QQ-plot. There aren't any outliers.

```
ggplot(pima1, aes(x=diabetes, y=pressure, fill=diabetes)) +
  geom_boxplot() +
  scale_y_log10() +
  scale_fill_brewer(palette="Set1", direction=-1) +
  labs(y="Blood pressure",
       title="Higher blood pressure associated with diabetes") +
  theme_minimal() + theme(plot.title = element_text(hjust = 0.5))
```



Hypothesis tests

Does diabetes affect blood pressure?

- H0: there is no relationship between blood pressure and diabetes
- H1: there is a relationship between blood pressure and diabetes

```
summary(fit1)
```

```
##
## Call:
## lm(formula = pressure ~ diabetes, data = pima1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.688  -7.413   0.587   7.312  36.587
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  69.4131     0.7158  96.977  < 2e-16 ***
## diabetespos   5.2744     1.2446   4.238 2.83e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 11.52 on 385 degrees of freedom
## Multiple R-squared:  0.04457,	Adjusted R-squared:  0.04209
## F-statistic: 17.96 on 1 and 385 DF,  p-value: 2.827e-05
```

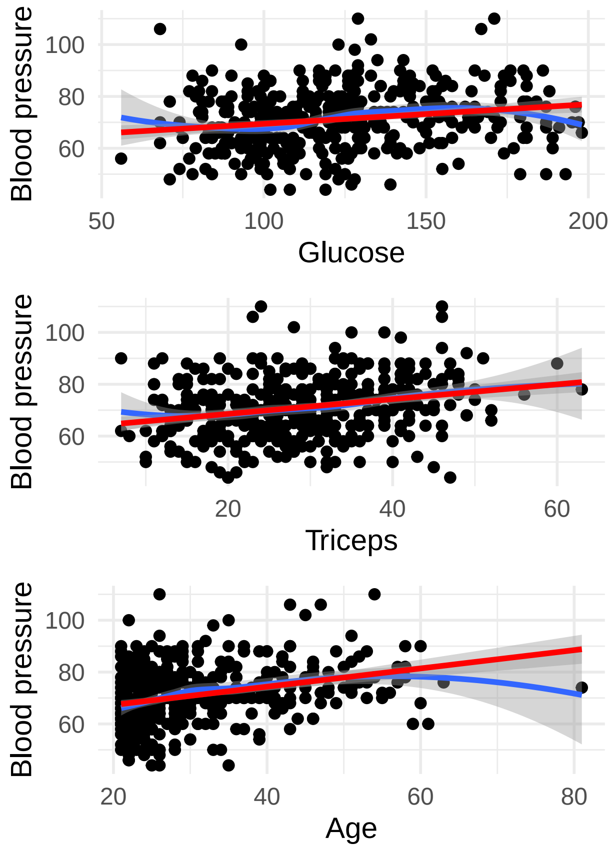
Explanation for Problem 3:

The variable is significant with p-value: 2.827e-05 & at alpha = 0.05 significance, we would reject H0 since high blood pressure is associated with diabetes.

Problem 4 ->

```r
g1 <- ggplot(pima1, aes(x=glucose, y=pressure)) +
        geom_point() + geom_smooth() + geom_smooth(method="lm", color="red") +
          labs(x="Glucose", y="Blood pressure") +
            theme_minimal()


g2 <- ggplot(pima1, aes(x=insulin, y=pressure)) +
        geom_point() + geom_smooth() + geom_smooth(method="lm", color="red") +
          labs(x="Insulin", y="Blood pressure") +
            theme_minimal()


g3 <- ggplot(pima1, aes(x=triceps, y=pressure)) +
        geom_point() + geom_smooth() + geom_smooth(method="lm", color="red") +
          labs(x="Triceps", y="Blood pressure") +
            theme_minimal()

g4 <- ggplot(pima1, aes(x=mass, y=pressure)) +
        geom_point() + geom_smooth() + geom_smooth(method="lm", color="red") +
          labs(x="Mass", y="Blood pressure") +
            theme_minimal()

g5 <- ggplot(pima1, aes(x=age, y=pressure)) +
        geom_point() + geom_smooth() + geom_smooth(method="lm", color="red") +
          labs(x="Age", y= "Blood pressure") +
             theme_minimal()

gridExtra::grid.arrange(g1, g2, g3, g4, g5, ncol = 2, nrow = 3)
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



From the scatterplot, there appears to be a positive linear association of mass, age & triceps with blood pressure, so we will use them in the model.

Possible covariates in Model 1: mass Possible covariates in Model 2: mass, age Possible covariates in Model 3: mass, age, triceps

```
#fit three models
model1 <- lm(pressure ~ diabetes + mass, data = pima1)
model2 <- lm(pressure ~ diabetes + mass + age, data = pima1)
model3 <- lm(pressure ~ diabetes + mass + age + triceps , data = pima1)
```

```
library(AICcmodavg)
```

```
## Warning: package 'AICcmodavg' was built under R version 4.2.2
```

```
#define list of models
models <- list(model1, model2, model3)

#specify model names
mod.names <- c('diab.mass', 'diab.mass.age', 'diab.mass.age.tri')

#calculate AIC of each model
aictab(cand.set = models, modnames = mod.names)
```

```
##
```

```
## Model selection based on AICc:
##
##                    K    AICc Delta_AICc AICcWt Cum.Wt       LL
## diab.mass.age      5 2922.60       0.00   0.69   0.69 -1456.22
## diab.mass.age.tri  6 2924.23       1.62   0.31   1.00 -1456.00
## diab.mass          4 2951.54      28.93   0.00   1.00 -1471.72
```

Explanation for Problem 4:

The model with the lowest AIC value is always listed first. From the output we can see that the following model having diabetes, mass & age have the lowest AIC value and is thus the best fitting model.

Problem 5 ->

```
fit2 <- lm(pressure ~ diabetes + mass + age, data=pima1)

summary(fit2)
```

```
##
## Call:
## lm(formula = pressure ~ diabetes + mass + age, data = pima1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.6467  -7.5024  -0.9172   7.5897  28.5663
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.14898    3.08592  13.658  < 2e-16 ***
## diabetespos  0.65829    1.24987   0.527    0.599
## mass         0.57781    0.07979   7.242 2.45e-12 ***
## age          0.31431    0.05563   5.650 3.13e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.48 on 383 degrees of freedom
## Multiple R-squared:  0.2139, Adjusted R-squared:  0.2077
## F-statistic: 34.73 on 3 and 383 DF,  p-value: < 2.2e-16
```

Hypothesis tests

Does diabetes affect blood pressure?

- H0: there is no relationship between blood pressure and diabetes
- H1: there is a relationship between blood pressure and diabetes

Explanation for Problem 5:

The explanatory variable diabetes with p-value: 0.599 & at alpha = 0.05 significance, we fail to reject the H0.

The results are different since after accounting for the effect of mass & age, diabetes variable becomes insignificant.

```r
#load the car library
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some
```

```r
#calculate the VIF for each predictor variable in the model
vif(fit2)
```

```
## diabetes     mass      age
## 1.219318 1.072902 1.142730
```

In this case, we may want to remove diabetes from the model because it has a high VIF value and it was not statistically significant at the 0.05 significance level.