Exploratory Data Analysis and Visualization of HbA1c Dataset in Jordan

Prepared by Mohammad Awad

Task

You are provided with a dataset (HbA1cDataset.csv). The dataset represents demographics and factors that may increase the risk of high blood pressure. The following list explains the variables in the dataset: 1- City: the residence of the participant. 2- Age: age of the participant in years. 3- Sex: Male or female. 4- Smoking: the participant's smoking status (Yes: Smoker, No: Non-smoker). 5- BMI: Body Mass Index. 6- HbA1c: a blood test that measures average blood sugar levels over the past three months. 7- Diabetic_status: represents the diabetic status of the participants a. HbA1c ≥ 6.5 indicates diabetes b. HbA1c ≥ 5.7 indicates pre-diabetes c. HbA1c < 5.7 Normal. 8- SBP: Systolic Blood pressure.

Part A

Calculate the correlation coefficient between Age and SBP. Explain the correlation coefficient. Use two methods to check the bivariate normal distribution condition before calculating the correlation.

```
library(pacman)
p_load(tidyverse, MASS, car, dplyr)

# Data cleaning
q2 <- read.csv("HbA1cDataset.csv")
glimpse(q2)
```

```
## Rows: 200
## Columns: 8
## $ City            <chr> "Amman", "Amman", "Amman", "Amman", "Amman", "Amman", ~
## $ Age             <int> 28, 62, 41, 72, 66, 38, 45, 29, 46, 36, 21, 80, 46, 55~
## $ Sex             <chr> "Male", "Male", "Female", "Male", "Male", "Female", "F~
## $ Smoking         <chr> "Yes", "No", "No", "Yes", "Yes", "No", "Yes", "No", "Y~
## $ BMI             <dbl> 22.6, 26.8, 25.0, 30.3, 31.9, 21.9, 26.8, 28.6, 32.4, ~
## $ HbA1C           <dbl> 5.8, 6.0, 5.8, 7.2, 6.4, 5.5, 6.4, 6.1, 7.8, 5.9, 5.8,~
## $ SBP             <int> 116, 115, 108, 144, 135, 114, 136, 119, 146, 124, 124,~
## $ Diabetic_status <chr> "Prediabetic", "Prediabetic", "Prediabetic", "Diabetic~
```

```
q2$City <- factor(q2$City)
q2$Sex <- factor(q2$Sex)
q2$Smoking <- factor(q2$Smoking)
q2$Diabetic_status <- factor(q2$Diabetic_status, ordered = TRUE, levels = c("Normal", "Prediabet
ic", "Diabetic"))

# Summarizing the data
summary(q2$BMI)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.80   24.60   28.00   28.33   31.12   45.20
```

```
summary(q2$HbA1C)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.000   5.800   6.250   7.005   7.900  13.600
```

```
summary(q2$SBP)
```

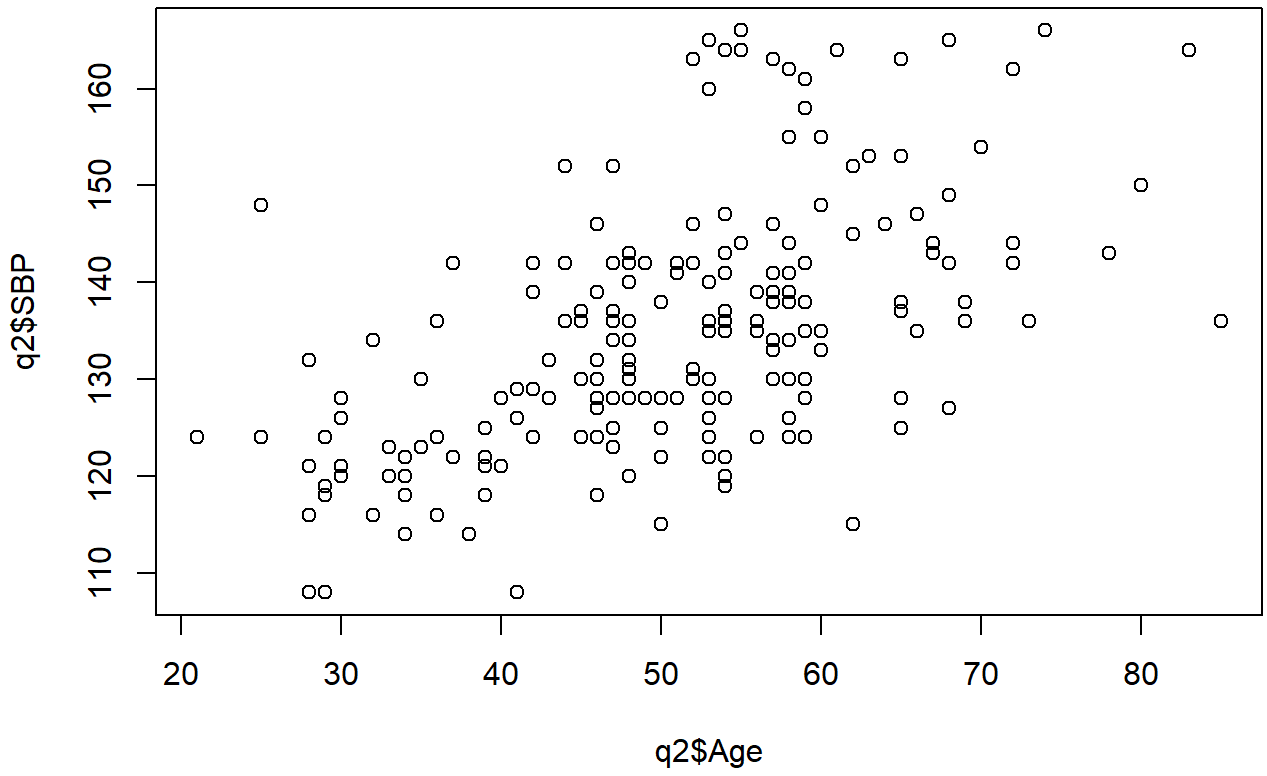
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   108.0   125.0   135.0   135.1   142.0   166.0
```

```
unique(q2$Diabetic_status)
```

```
## [1] Prediabetic Diabetic    Normal
## Levels: Normal < Prediabetic < Diabetic
```

Checking bivariate normality Age vs SBP

```
plot(q2$Age, q2$SBP)
```

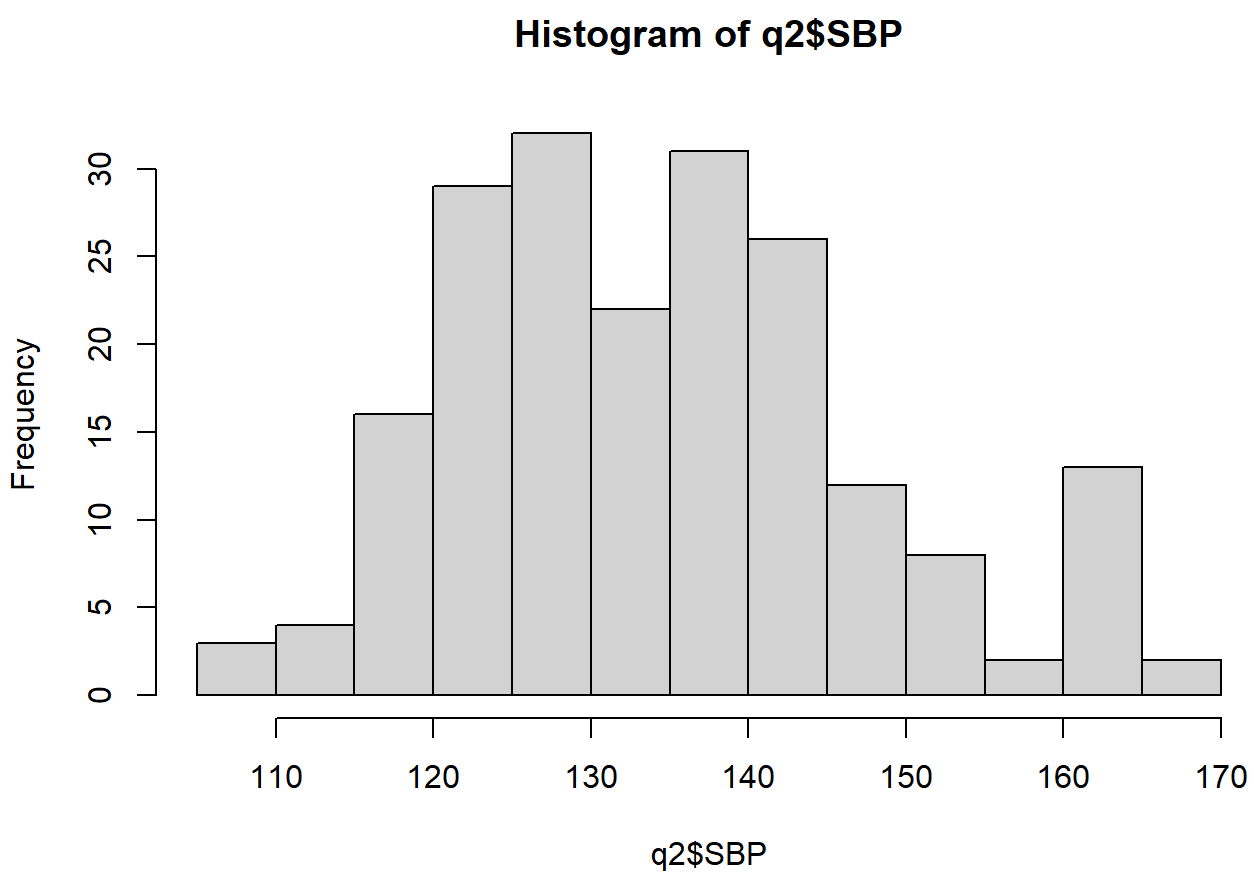


Age is skewed to the right

```
hist(q2$Age)
```

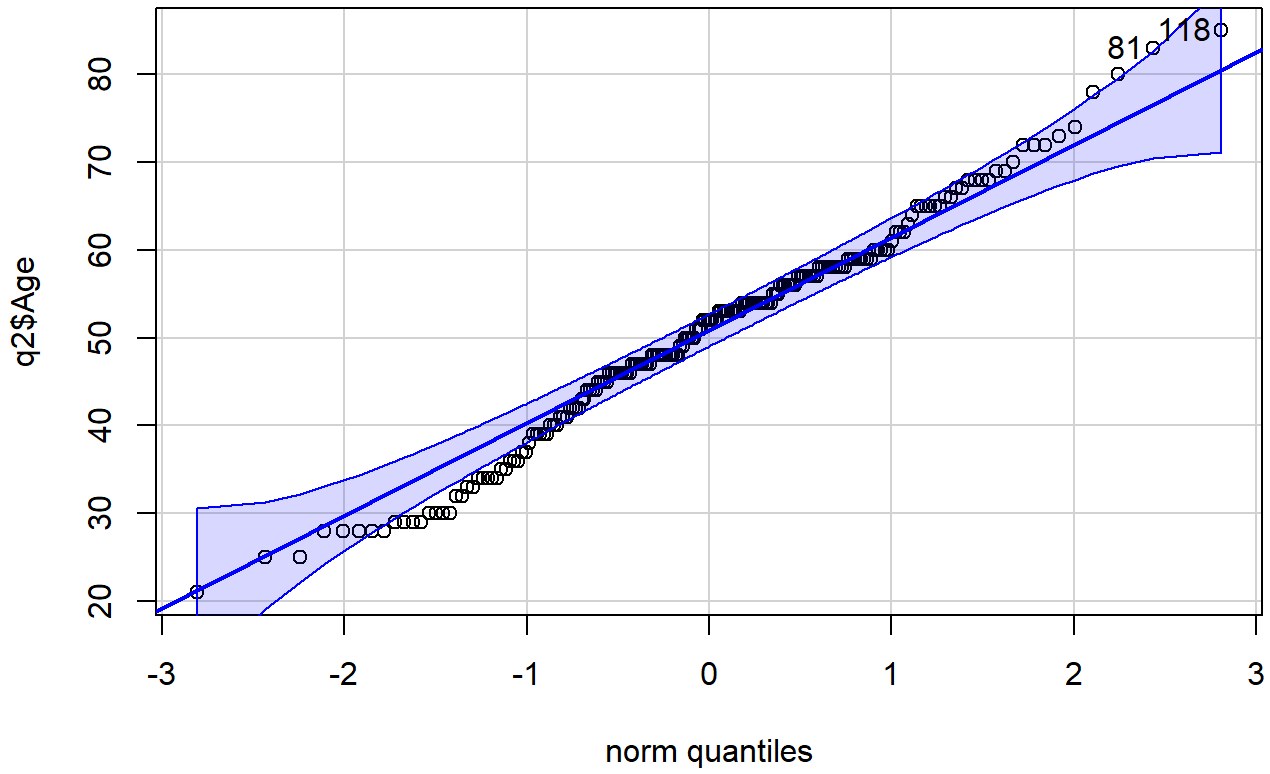
SBP distribution

```
hist(q2$SBP)
```



Q-Q Plot for Age

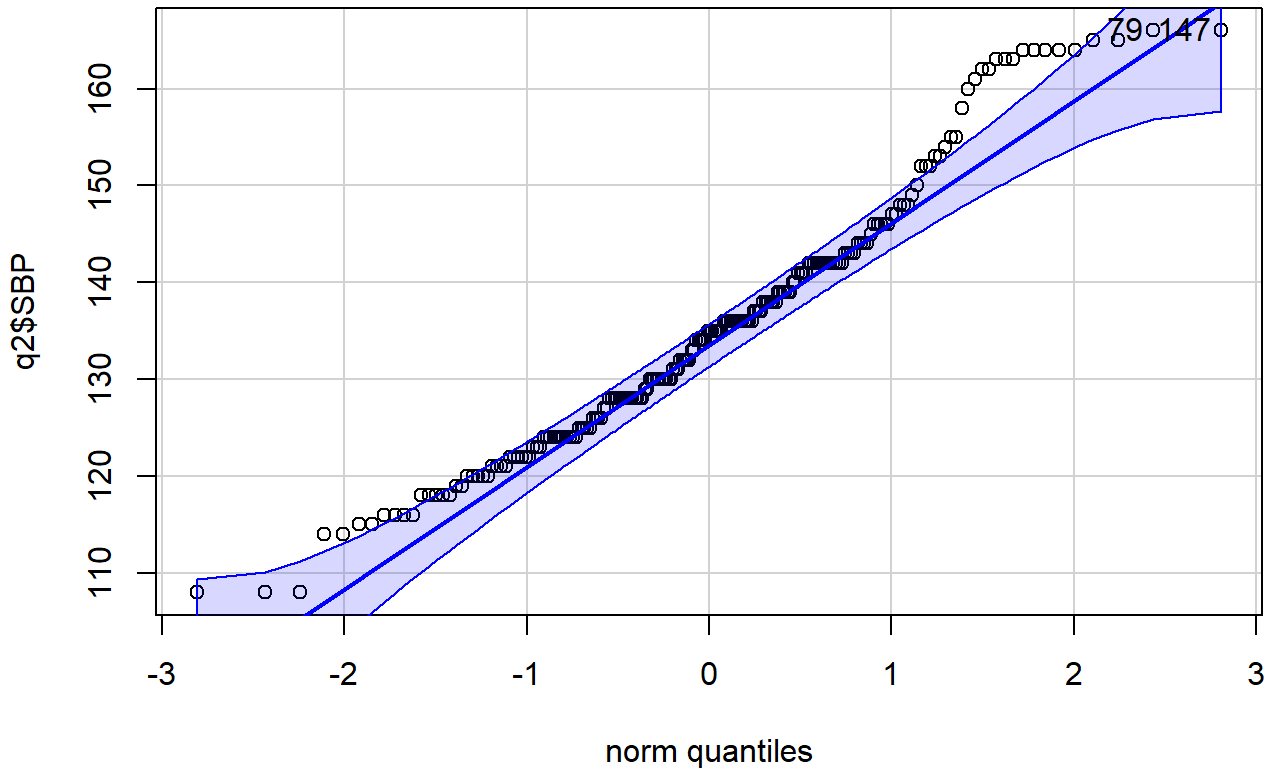
```
qqPlot(q2$Age)
```




```
## [1] 118  81
```

Q-Q Plot for SBP

```
qqPlot(q2$SBP)
```




```
## [1]  79 147
```

Correlation Coefficient

The correlation coefficient should not be used as the variables appear to be non-normal, but the values are calculated below:

```
cor(q2$Age, q2$SBP)
```

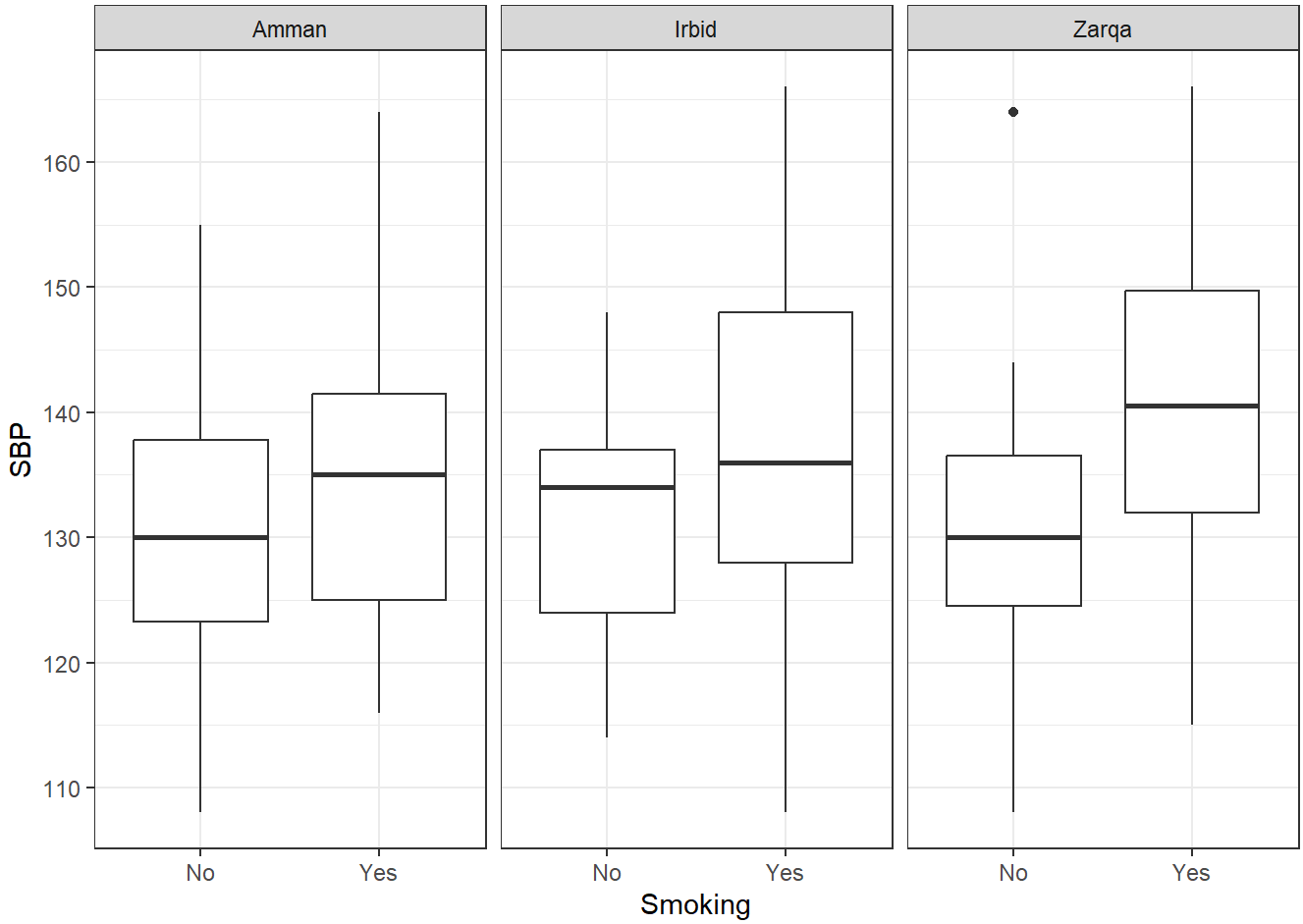
```
## [1] 0.5627083
```

`=0.5627083` shows a positive correlation, meaning that Age can be used to explain variability in SBP, but as they're not bivariate normal, it is not recommended to use this correlation value.

Part B

Create a visualization that shows the relationship between SBP, Smoking, and the City of residence of the participants. Interpret the visualization you created in no more than 50 words. The visualization should adhere to the best practices you learned.

```
q2 %>% ggplot(aes(x = Smoking, y = SBP)) +
  geom_boxplot() +
  facet_wrap(vars(City)) +
  theme_bw()
```



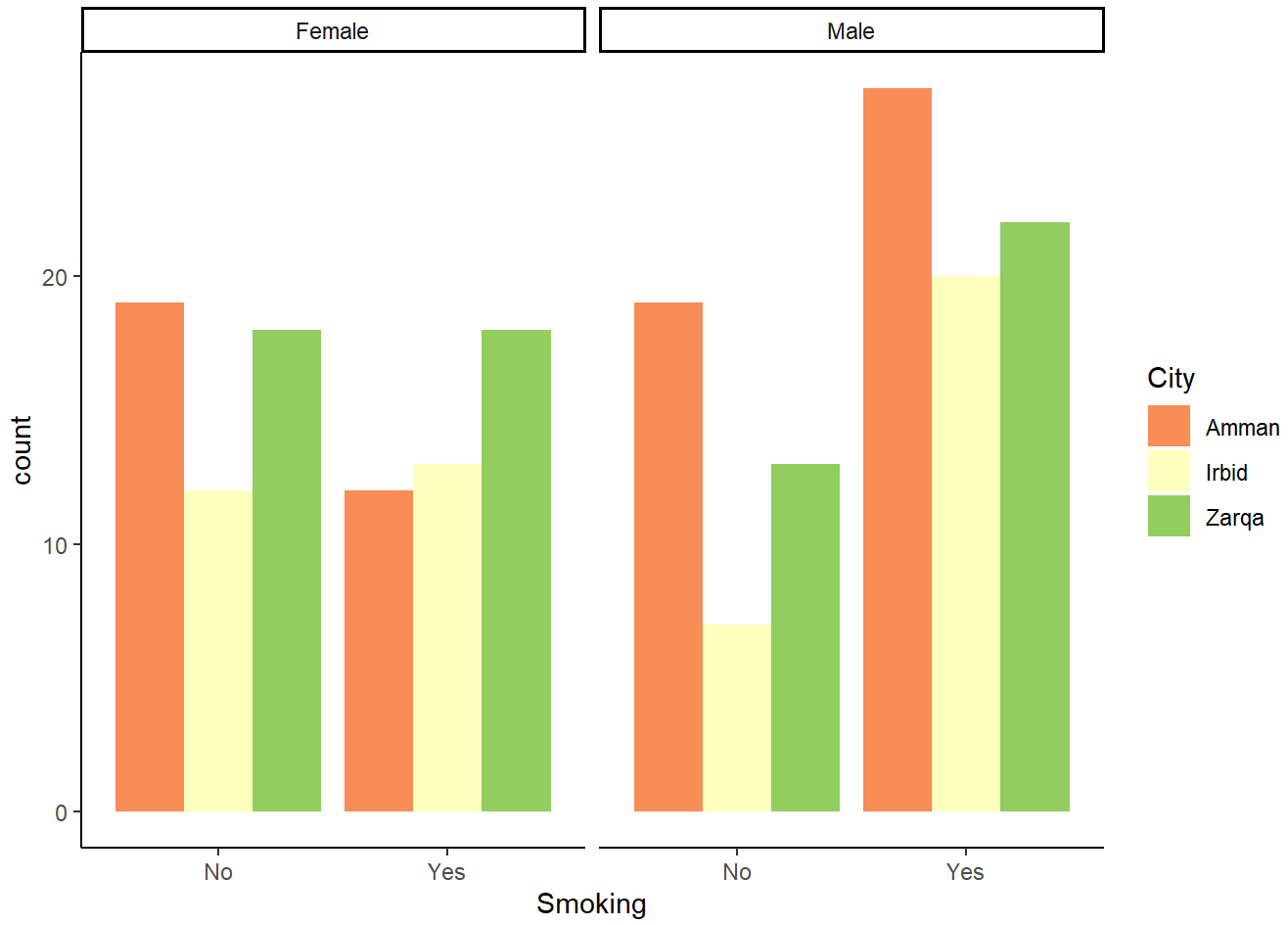
Interpretation:

1. City of Zarqa showed the highest difference between the median of smokers and non-smokers, with one potential outlier for non-smokers on the upper end.
2. City of Irbid showed the widest range of SBP values for smokers, while the median appears close for both smokers and non-smokers.
3. City of Amman appears to have the least variance between smokers and non-smokers, but non-smokers showed a wider range of values, especially on the upper end compared to the other two cities. All cities showed a higher median of SBP for smokers vs non-smokers.

Part C

Create a visualization that shows the relationship between smoking status, sex, and the city of residence. The visualization should allow for a direct comparison. Interpret the plot by answering the following points: a. Which city has the highest percentage of smokers among males? b. Which city has the highest percentage of smokers among females?

```
q2 %>% ggplot(aes(x = Smoking, fill = City)) +
  geom_bar(position = "dodge") +
  facet_wrap(vars(Sex)) +
  theme_classic() +
  scale_fill_brewer(palette = "RdYlGn")
```



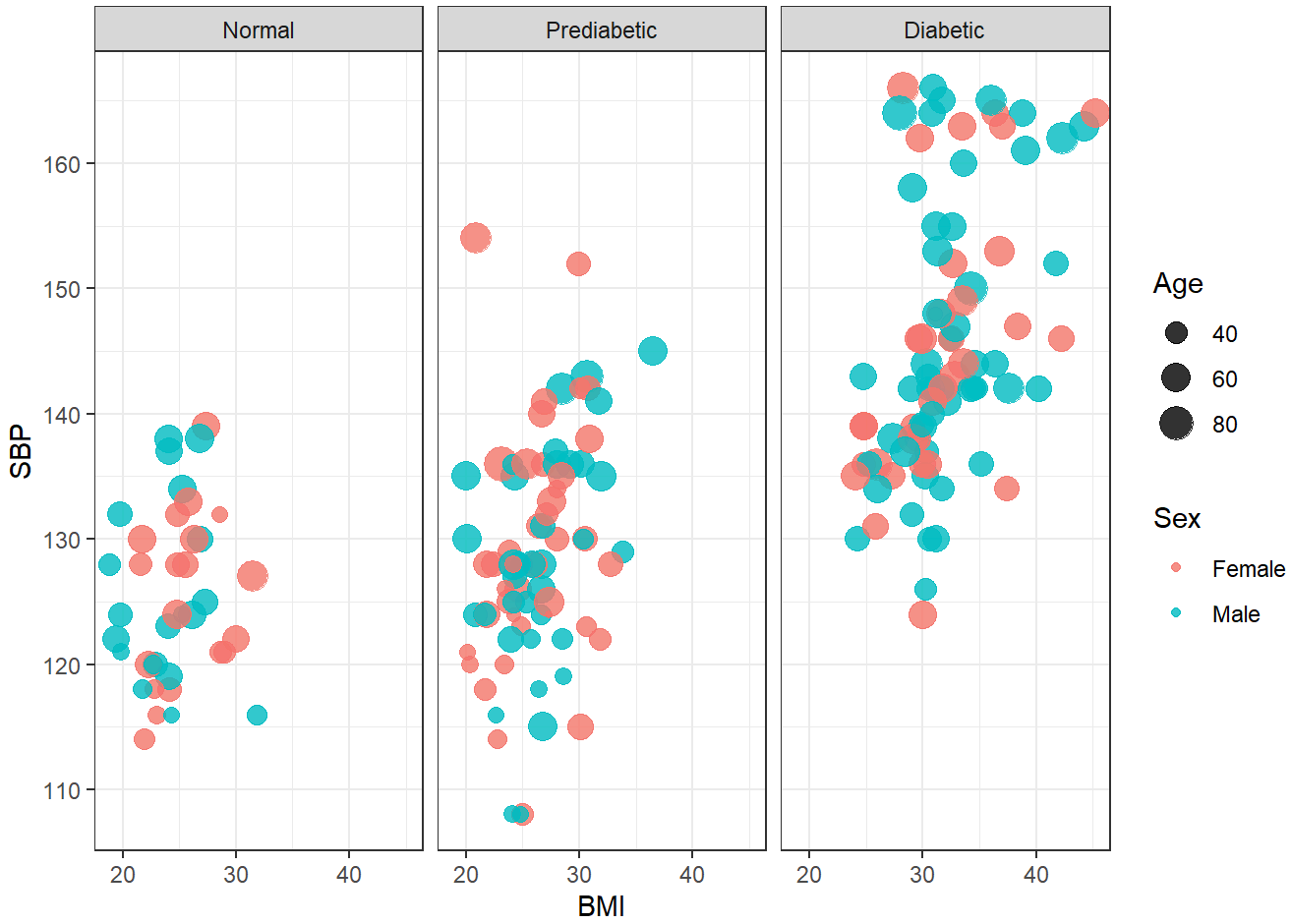
Interpretation:

a. The city of Amman has the highest count of smokers among males.
b. The city of Zarqa has the highest count of smokers among females.

Part D

Create a visualization that shows the effect of BMI on SBP, adjusted for age, sex, and diabetic status. Explain your visualization in no more than 50 words.

```
q2 %>% ggplot(aes(x = BMI, y = SBP, size = Age, color = Sex)) +
  geom_point(alpha = 0.8) +
  facet_wrap(vars(Diabetic_status)) +
  theme_bw()
```



Interpretation:

For both males and females, an increase in BMI corresponds with an increase in SBP in diabetics but not as much in normal and prediabetic individuals. Normal and prediabetics have a higher proportion of younger people, while diabetics consist more of older people.