COMP 1433: Introduction to Data Analytics &
COMP 1003: Statistical Tools and Applications

Lecture 7 – Programming with R: Statistics

Dr. Jing Li
Department of Computing
The Hong Kong Polytechnic University

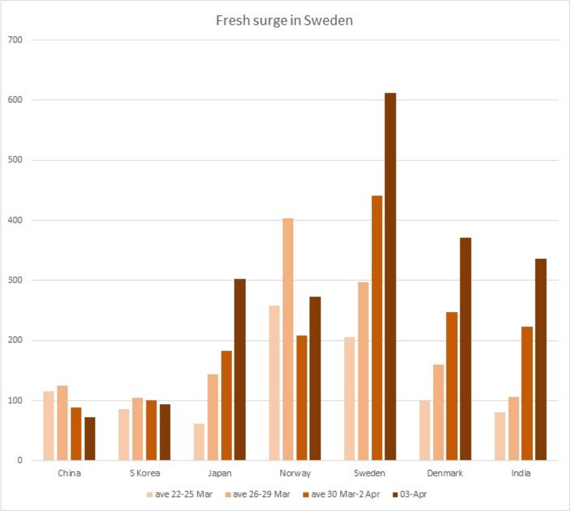
1&3 Mar 2022

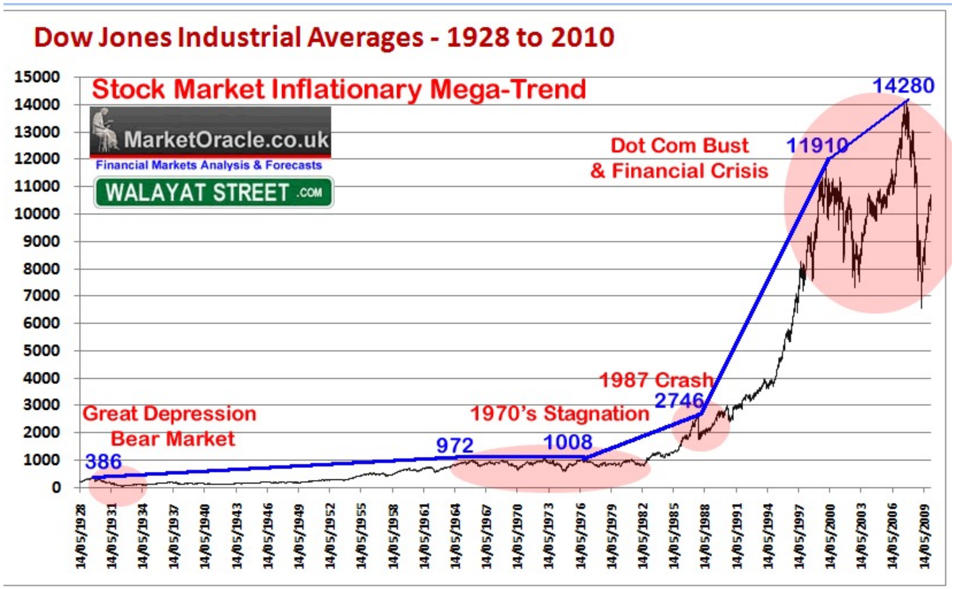
Keypoints Recap (R Programing)

- Different types of data representation:
  - *Vectors*, *Matrices*, *Arrays*, *Dataframes*, *Lists*, etc.
- Import data (from texts), viewing data, and exporting results (to texts).
- Data Manipulation
  - Control Structure (e.g., for loop, if condition, etc.)
  - Arithmetic and Logical Operations
  - **Built-in Functions**: numeric, character, probability, statistics, etc.

Graphs vs. Statistics

- Numbers are boring while graphs are straightforward
- Visualize the data
- Helpful to analyze the data statistics





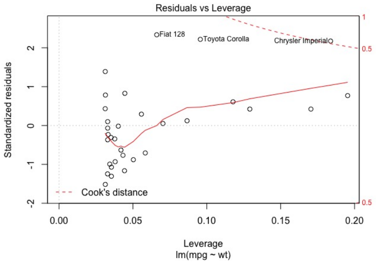
ggplot2 Package

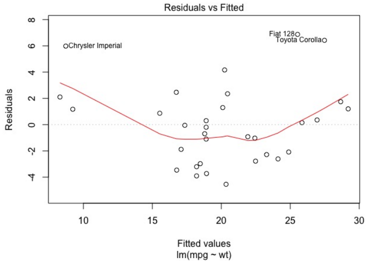
- R has built-in functions for charts and graphs (base graphics), such as *plot().*

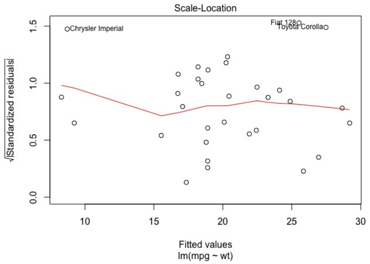
- The R package *ggplot2* extend the charting and graphing functions.

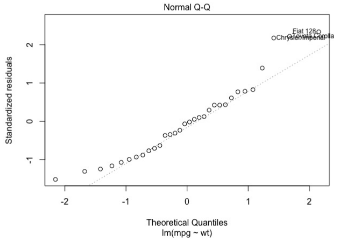
    install.packages("ggplot2")

    library("ggplot2")









Roadmap

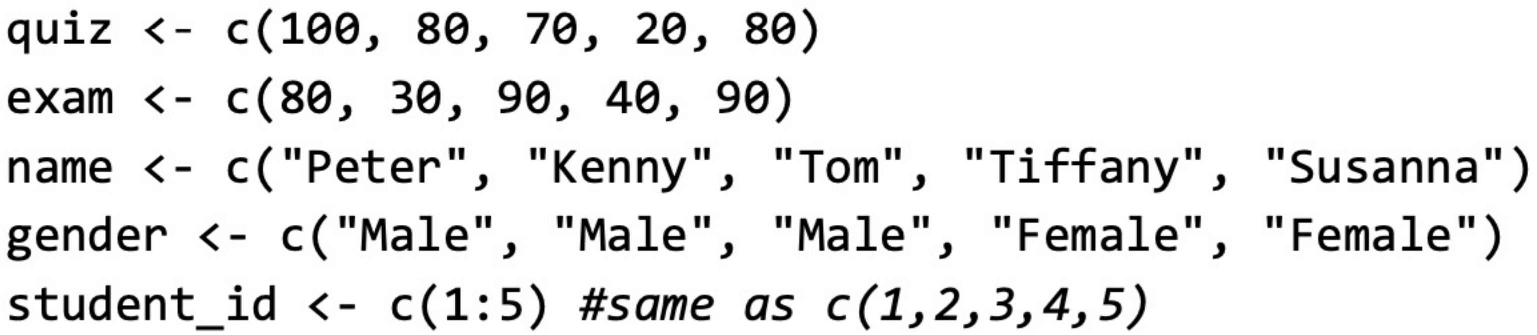
- Barplot
- Histograms
- Scatterplot
- **Example**: *Big Mart Sales Datasets*

Roadmap

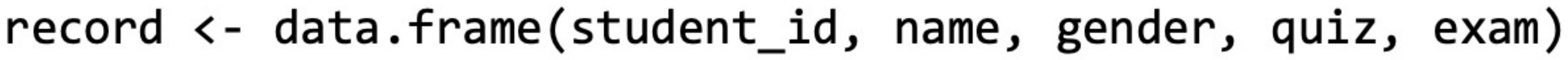
- **Barplot**
- Histograms
- Scatterplot
- **Example**: *Big Mart Sales Datasets*

Barplot

- Input Data

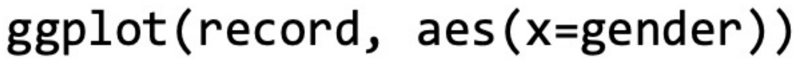


```
quiz <- c(100, 80, 70, 20, 80)
exam <- c(80, 30, 90, 40, 90)
name <- c("Peter", "Kenny", "Tom", "Tiffany", "Susanna")
gender <- c("Male", "Male", "Male", "Female", "Female")
student_id <- c(1:5) #same as c(1,2,3,4,5)
```



```
record <- data.frame(student_id, name, gender, quiz, exam)
```

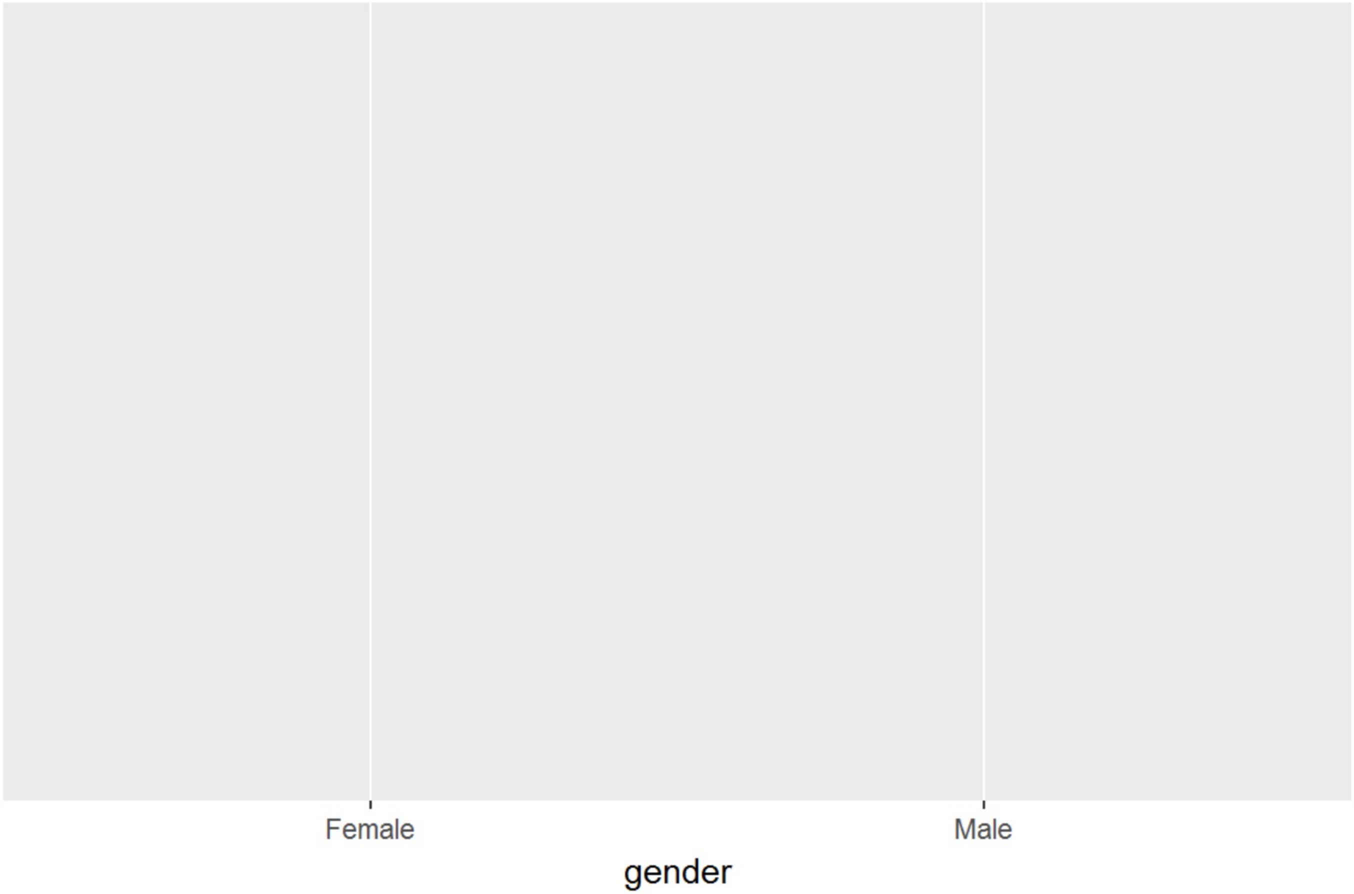
- Define a chart:



```
ggplot(record, aes(x=gender))
```

Barplot

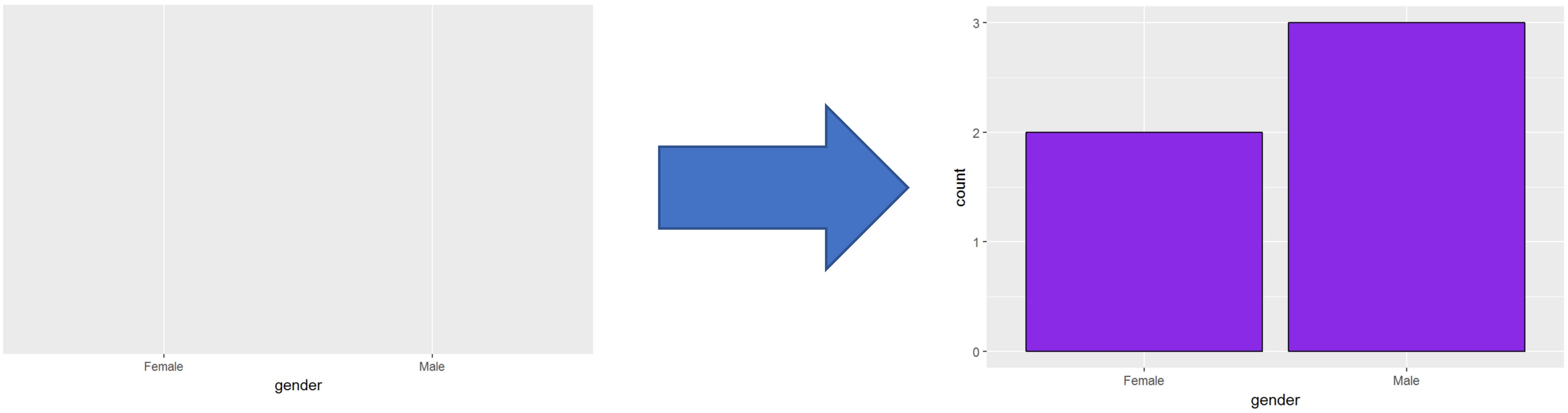
- Define a chart: `ggplot(record, aes(x=gender))`



Barplot

- The chart can then be enhanced step by step.
  - E.g. creates a bar chart and fill it with blueviolet color and black border.

```
chart <- ggplot(record,aes(x=gender))
bars <- geom_bar(fill="blueviolet", color="black")
chart+bars
```



Barplot

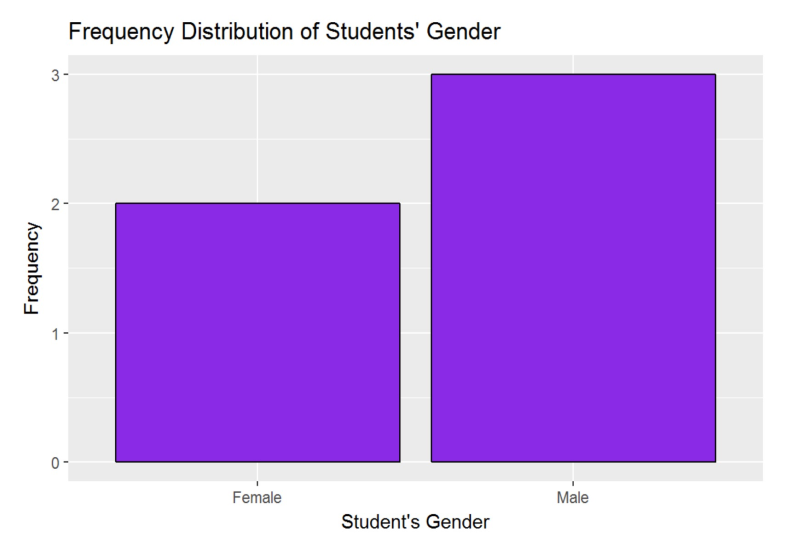
- Add additional commands to specify the chart/axis titles.
  - ***ggtitle***: chart title; ***xlab,ylab***: Label for x-axis/y-axis

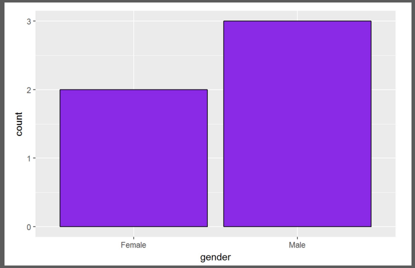
```
xlabel <- xlab("Student's
Gender")

ylabel <- ylab("Frequency")

title <- ggtitle("Frequency
Distribution of Students' Gender
")

chart+bars+xlabel+ylabel+title
```

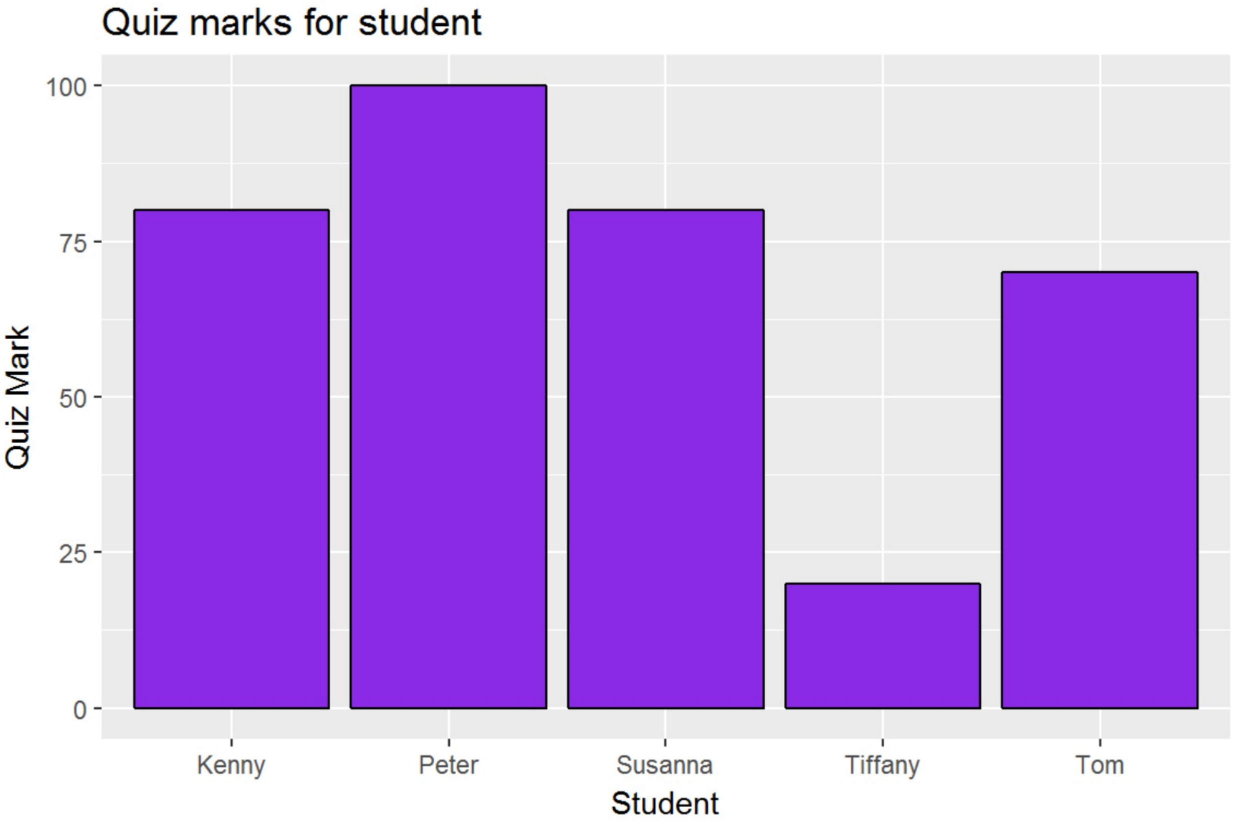




Barplot

Identity: Make the heights of the bars to represent values in the data.

```
ggplot(record, aes(x=name, y=quiz))  +
  geom_bar(fill="blueviolet", color="black", stat="identity") +
  xlab("Student")+ ylab("Quiz Mark") + ggtitle("Quiz marks for student")
```



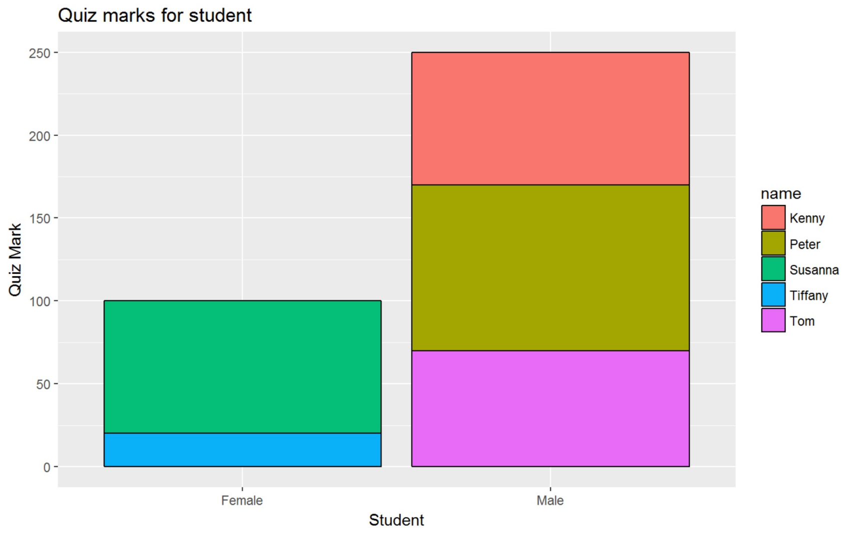
What if we want to color the bars with students' genders?

Barplot

- First group by gender and create individual bars for each student.

Within each gender, we breakdown the data into different students and fill the bars with different colors.

```
ggplot(record, aes(x=gender, y=quiz, fill=name))  +
  geom_bar(color="black", stat="identity") +
  xlab("Student")+ ylab("Quiz Mark") + ggtitle("Quiz marks for student")
```



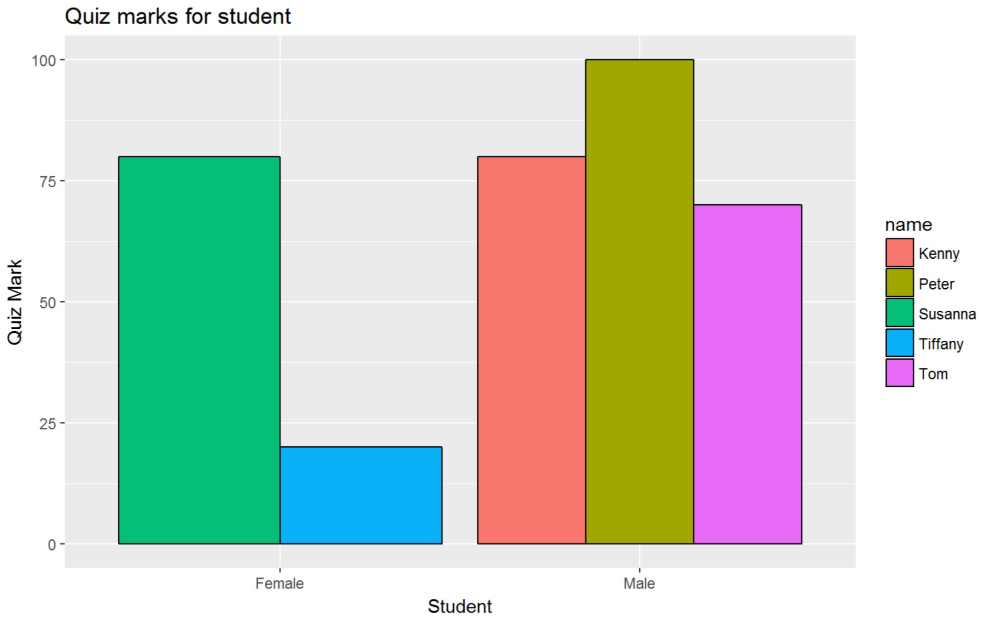
A stacked bar plot is created by default. How about interleaved bars?

Barplot

- First group by gender and create individual bars for each student.

To create interleaved bars!

```
ggplot(record, aes(x=gender, y=quiz, fill=name))  +
  geom_bar(color="black", stat="identity", position="dodge") +
  xlab("Student")+ ylab("Quiz Mark") +
  ggtitle("Quiz marks for student")
```



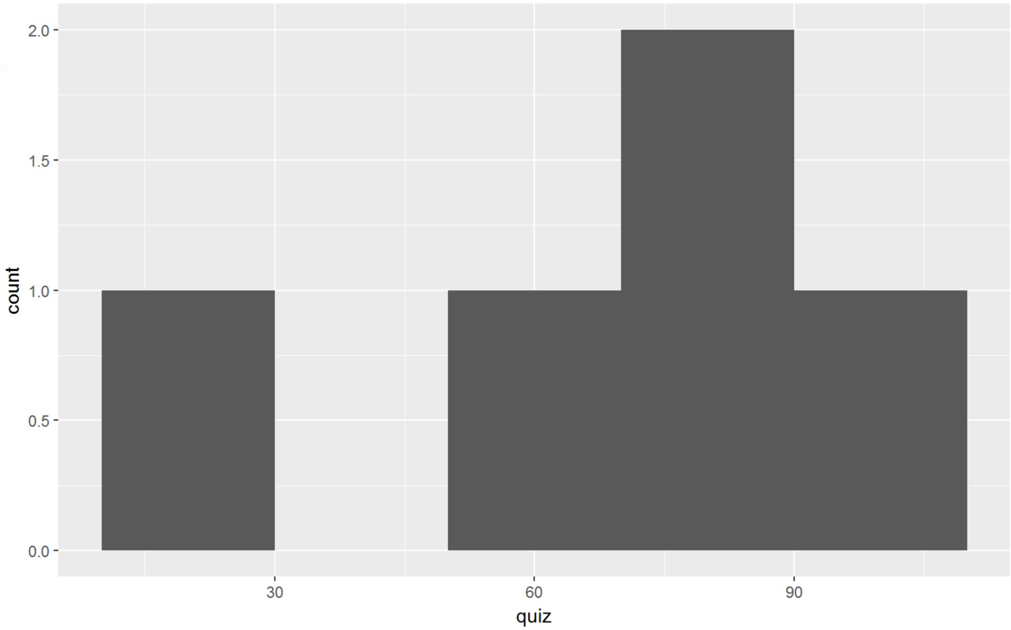
Roadmap

- Barplot
- **Histograms**
- Scatterplot
- **Example**: *Big Mart Sales Datasets*

Histograms

- A histogram consists of parallel vertical bars that graphically shows the frequency distribution of a quantitative variable (e.g. quiz, exam).
  - We can use them to visualize data distributions!

```
ggplot(record, aes(x=quiz))  +
  geom_histogram(binwidth=20)
```



You may further define a color palette for plotting charts.
Explore by yourselves!

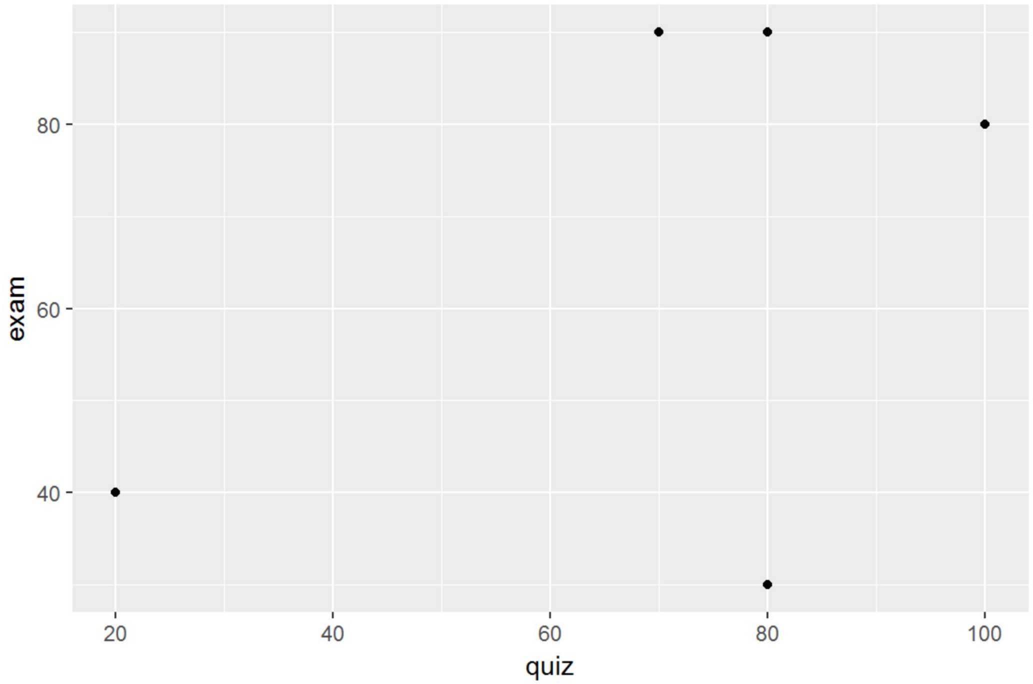
Roadmap

- Barplot
- Histograms
- **Scatterplot**
- **Example**: *Big Mart Sales Datasets*

Scatter Plot

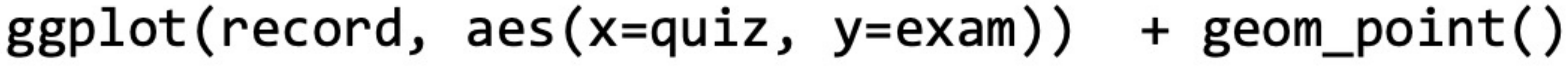
- Add the points using a geom layer called geom_point with the ‘+’ operator.

```
ggplot(record, aes(x=quiz, y=exam))  + geom_point()
```

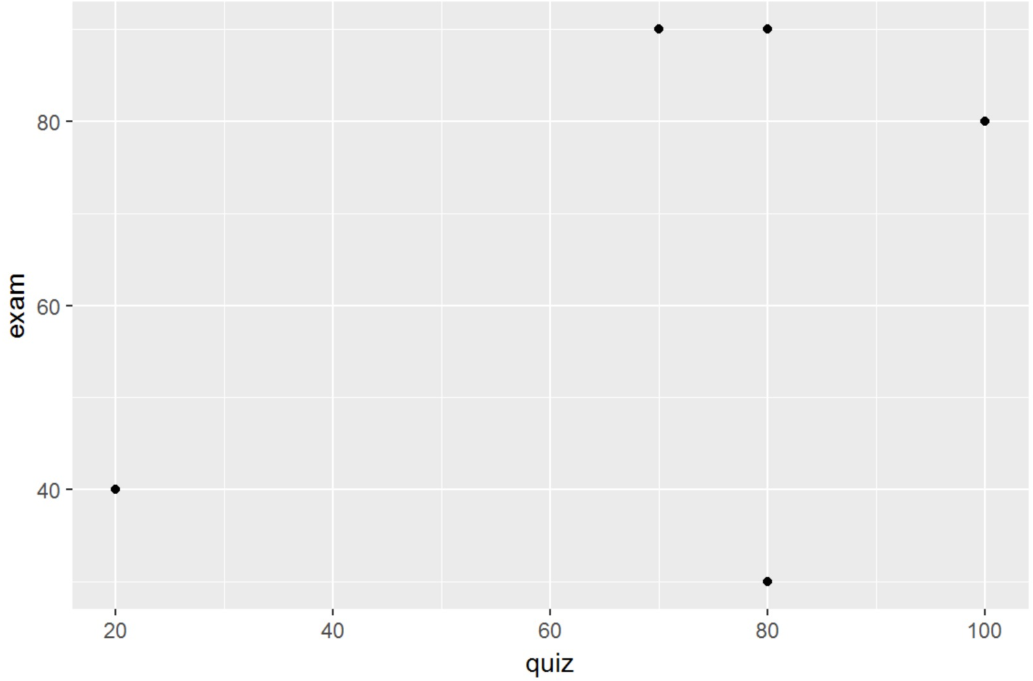


Scatter Plot

- Add the points using a geom layer called geom_point with the ‘+’ operator.

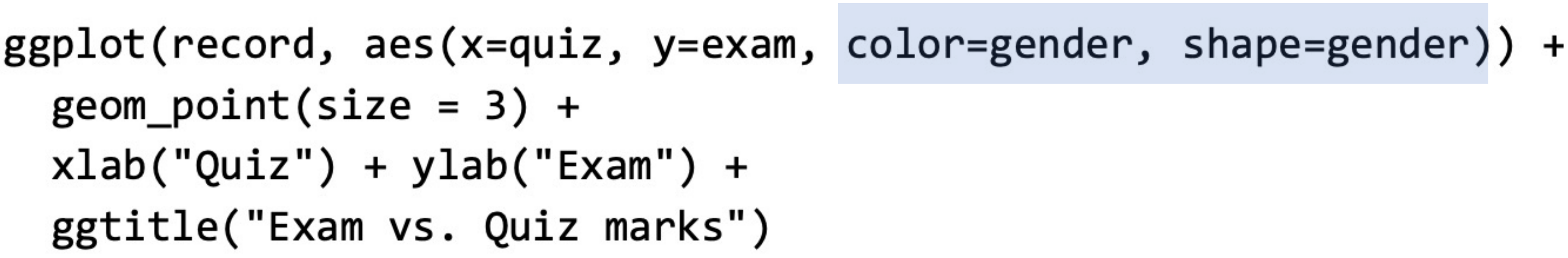


```
ggplot(record, aes(x=quiz, y=exam))  + geom_point()
```

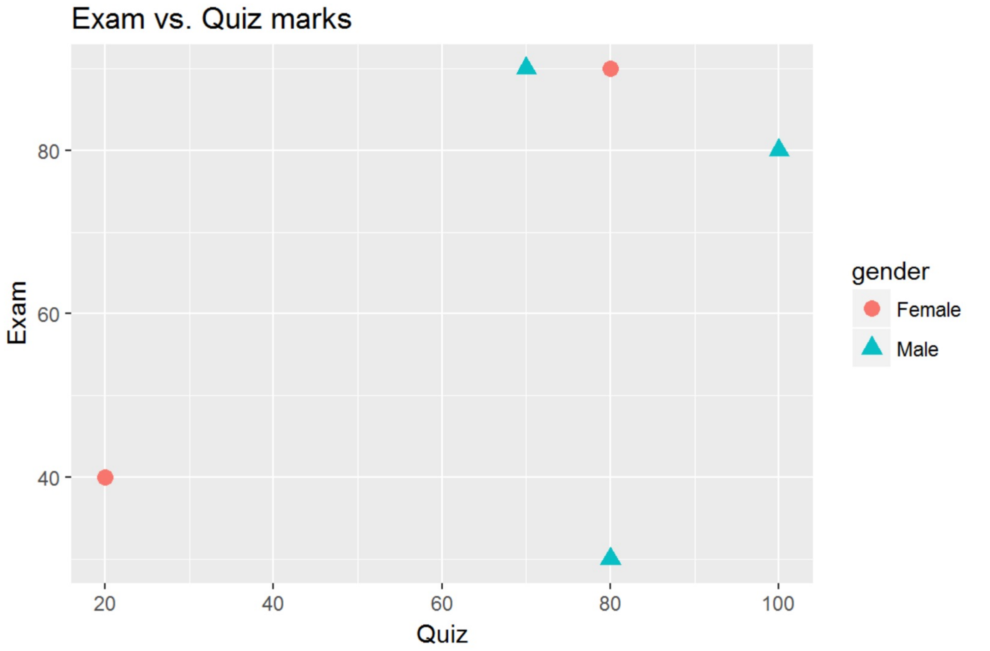


Scatter Plot

- customize the point size and color of the points, add the axis titles and chart title.



```
ggplot(record, aes(x=quiz, y=exam, color=gender, shape=gender)) +
  geom_point(size = 3) +
  xlab("Quiz") + ylab("Exam") +
  ggtitle("Exam vs. Quiz marks")
```

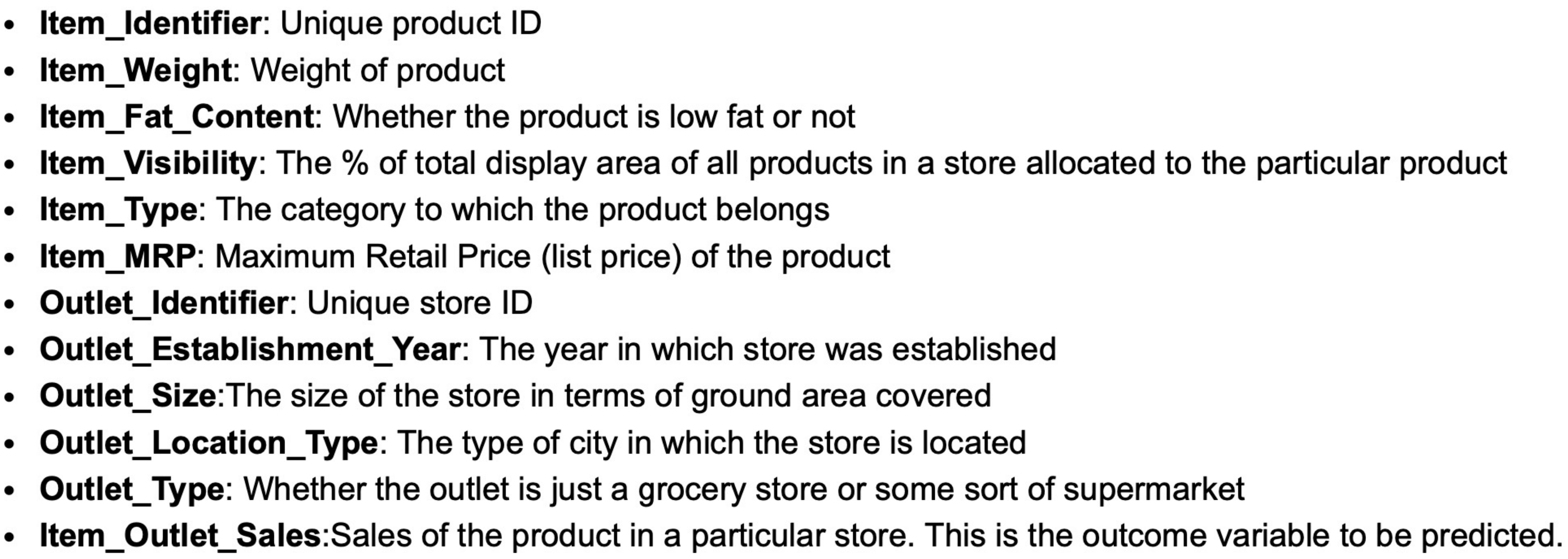


Roadmap

- Barplot
- Histograms
- Scatterplot
- **Example: *Big Mart Sales Datasets***

Example: Big Mart Sales Dataset

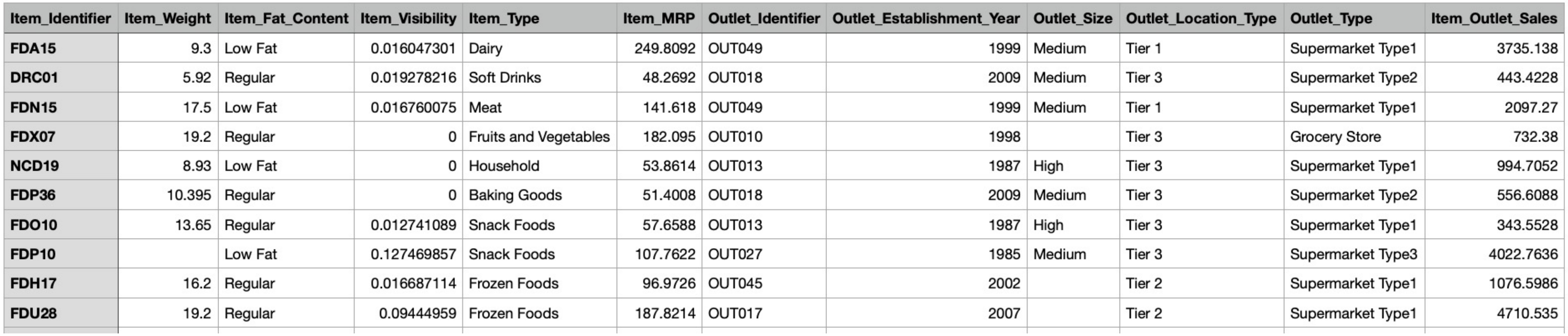
- Data Description:
  - The data scientists at Big Mart collected 2013 sales data for 1559 products across 10 stores in different cities.



- **Item_Identifier**: Unique product ID
- **Item_Weight**: Weight of product
- **Item_Fat_Content**: Whether the product is low fat or not
- **Item_Visibility**: The % of total display area of all products in a store allocated to the particular product
- **Item_Type**: The category to which the product belongs
- **Item_MRP**: Maximum Retail Price (list price) of the product
- **Outlet_Identifier**: Unique store ID
- **Outlet_Establishment_Year**: The year in which store was established
- **Outlet_Size**:The size of the store in terms of ground area covered
- **Outlet_Location_Type**: The type of city in which the store is located
- **Outlet_Type**: Whether the outlet is just a grocery store or some sort of supermarket
- **Item_Outlet_Sales**:Sales of the product in a particular store. This is the outcome variable to be predicted.

Example: Big Mart Sales Dataset

- Data Description:
  - The data scientists at Big Mart collected 2013 sales data for 1559 products across 10 stores in different cities.



Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outlet_Type	Item_Outlet_Sales
FDA15	9.3	Low Fat	0.016047301	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermarket Type1	3735.138
DRC01	5.92	Regular	0.019278216	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Supermarket Type2	443.4228
FDN15	17.5	Low Fat	0.016760075	Meat	141.618	OUT049	1999	Medium	Tier 1	Supermarket Type1	2097.27
FDX07	19.2	Regular	0	Fruits and Vegetables	182.095	OUT010	1998		Tier 3	Grocery Store	732.38
NCD19	8.93	Low Fat	0	Household	53.8614	OUT013	1987	High	Tier 3	Supermarket Type1	994.7052
FDP36	10.395	Regular	0	Baking Goods	51.4008	OUT018	2009	Medium	Tier 3	Supermarket Type2	556.6088
FDO10	13.65	Regular	0.012741089	Snack Foods	57.6588	OUT013	1987	High	Tier 3	Supermarket Type1	343.5528
FDP10		Low Fat	0.127469857	Snack Foods	107.7622	OUT027	1985	Medium	Tier 3	Supermarket Type3	4022.7636
FDH17	16.2	Regular	0.016687114	Frozen Foods	96.9726	OUT045	2002		Tier 2	Supermarket Type1	1076.5986
FDU28	19.2	Regular	0.09444959	Frozen Foods	187.8214	OUT017	2007		Tier 2	Supermarket Type1	4710.535

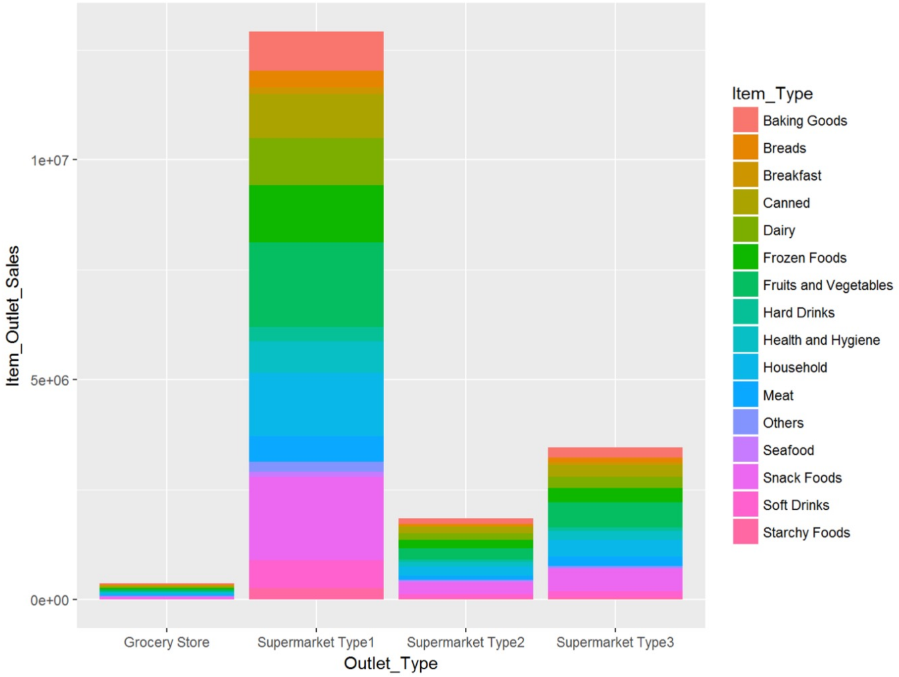
……

8523 records altogether

Example: Barplot for Sales

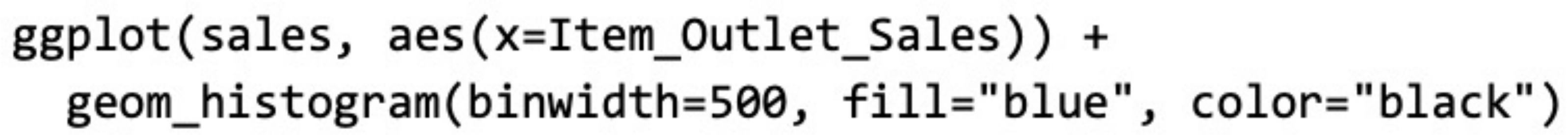
- Create a bar plot to show the total sales for different outlet types and add *fill=Item_Type* to ggplot to distinguish varying item types.

```
ggplot(sales, aes(x=Outlet_Type, y=Item_Outlet_Sales, fill=Item_Type)) +   geom_bar( stat="identity")
```

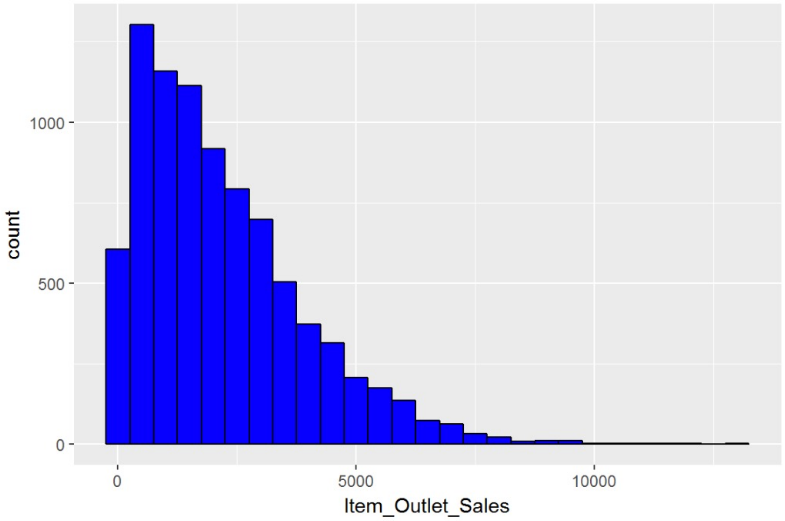


Example: Histograms for Sales

- Create a histogram to show the distribution of item outlet sales and fill the bars with *blue color* and *black border*.

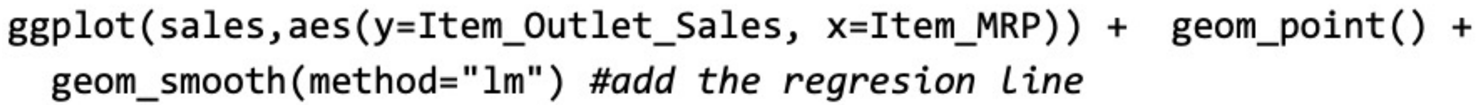


```
ggplot(sales, aes(x=Item_Outlet_Sales)) +
  geom_histogram(binwidth=500, fill="blue", color="black")
```

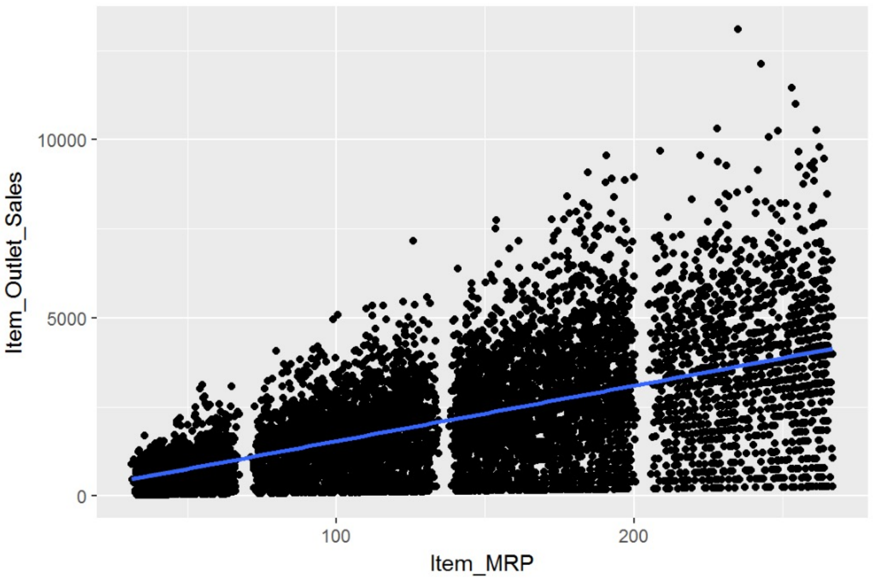


Example: Scatterplots for Sales

- Create a scatter plot to show the relationship between item retail price (x-axis) and item outlet sales (y-axis).



```
ggplot(sales,aes(y=Item_Outlet_Sales, x=Item_MRP)) +  geom_point() +
  geom_smooth(method="lm") #add the regresion line
```

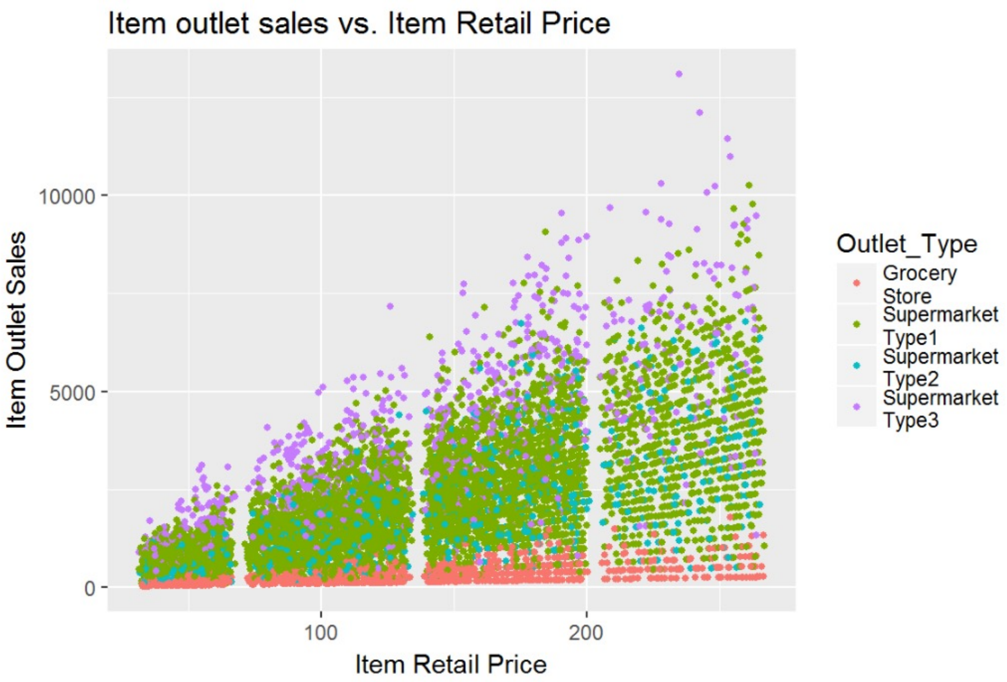


Regression Line

Example: Scatterplots for Sales

- Use different colors and shapes for data points with different outlet types.

```
ggplot(sales, aes(x=Item_MRP, y=Item_Outlet_Sales, color=Outlet_Type)) +  geom_point(size = 1) +
  xlab("Item Retail Price") + ylab("Item Outlet Sales") +
  ggtitle("Item outlet sales vs. Item Retail Price")
```

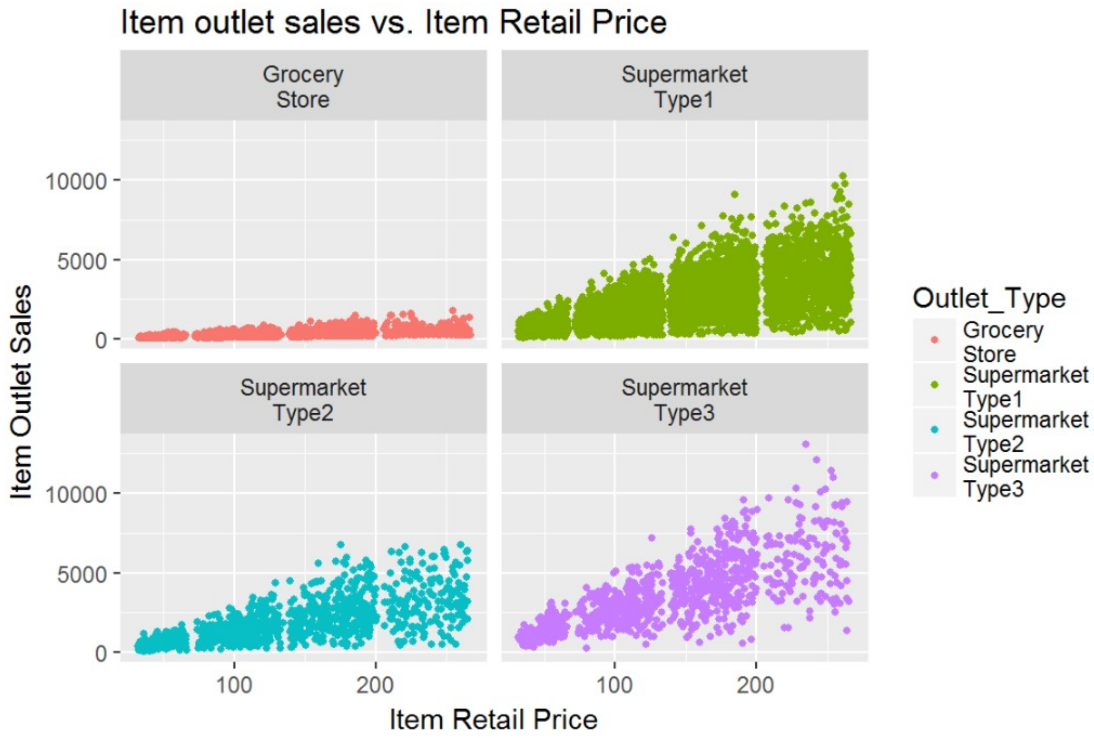


What can you find from the graph?

Example: Scatterplots for Sales

- Create a facet graph to create separate scatterplots for each outlet type.

```
ggplot(sales, aes(x=Item_MRP, y=Item_Outlet_Sales, color=Outlet_Type)) +  geom_point(size = 1) +
  xlab("Item Retail Price") + ylab("Item Outlet Sales") +
  ggtitle("Item outlet sales vs. Item Retail Price") +
  facet_wrap(~Outlet_Type)
```



A slide to take away

- How to draw barplot, histograms, and scatter plots for data?
- How to customize the graphs, e.g., the x and y axis, colors and the size of the bars and plots, etc.