Session 1

49-781 Data Analytics for Product Managers

Spring 2018

Session 1

Introduction

49-781 Data Analytics for Product Managers

Spring 2018

About me

Karimulla Shaikh (karim@cmu.edu)

Academic: IIT-Madras (1984-1991) & CMU (1991-1996)

Startup: Product Development at Several Startups (Production Information Management, RFID/Retail Analytics, Multimedia Production Platform, Language Translation, Customer-Relationship Intelligence, Software-Defined power)

Public Companies: Led Product Development at Two Public Companies (Medical Marketplace, Language Translation)

Interests: Product Management & Engineering; Products for Enterprises and Professionals; Traditional and Data-Driven; Focus on Cloud

About the course

- For product managers interested in delivering data products
  - Provides basic insights into the process of building data products (this is a mini course after all)
- Introduces and defines data products
  - We will distinguish them from traditional software products
- Introduces data analysis through quantitative methods
  - These are the most basic methods but provide the end-to-end experience of a typical product
- Gives hands-on experience through the development of a data product
  - Team project that goes from data to an end project
- Requires good understanding of Python
  - Prior experience with Python's data packages is desired but not required (You will have to learn them during the course though)

Resources

- There is no TA for this course, but you can reach out to me with any help you need
  - I prefer that you reach me on Slack. Use the channel *#instructor-help*
- I will provide guidance primarily in terms of pointing you to resources on the web. But I will not help you debug your code!
  - You can use Slack to help each other out. Use the channel *#student-discussion* so others may benefit from your experience.

Course Track 1

- Understand software development life cycle of a data product
  - You will do this through a team project that you will execute during the course
  - Project will be introduced in session 2 and will complete at the last session
  - You will do a hands-on data analytics on a relevant large data set for the project
- Track Components
  - Introduction to Data Products
  - Explore Existing Data Products
  - Dataset Exploration and Visualization
  - Data Product Development
  - Validation of Data Products

Course Track 2

- Technical introduction to underlying principles to build data products
  - Underlying statistical principles behind data analytics
  - Lecture and lab sessions to clarify and reinforce the concepts
  - Hands-on exercises on a variety of techniques for a number of data sets
- Track Components
  - Elementary Statistics
  - Linear Regression
  - Logistic Regression
  - Cross-validation
  - Non-linear regression
  - ‘Clustering Techniques

Tools & Technologies

- Python
  - Core language to be used in all labs, homework and projects
- NumPy
  - Numerical calculations using multidimensional arrays and matrices
- Pandas
  - Data munging and analysis. Dealing with missing and bad data.
- Matplotlib
  - Data visualization
- Statsmodels
  - Statistical models and tests.

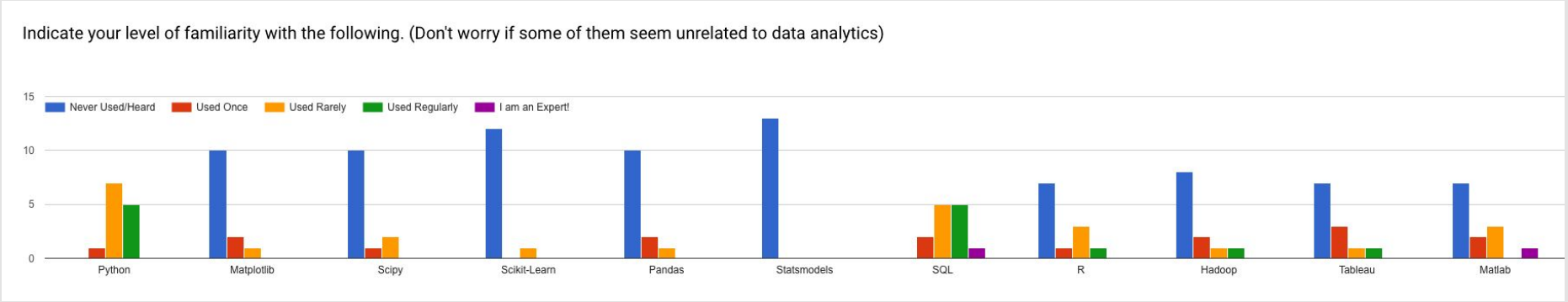
General Course Information

- Expected Course Load ~24 hours per week
- Tracks of Work: Lectures, Lab Activities, Quizzes, Homework and Project
- Grading

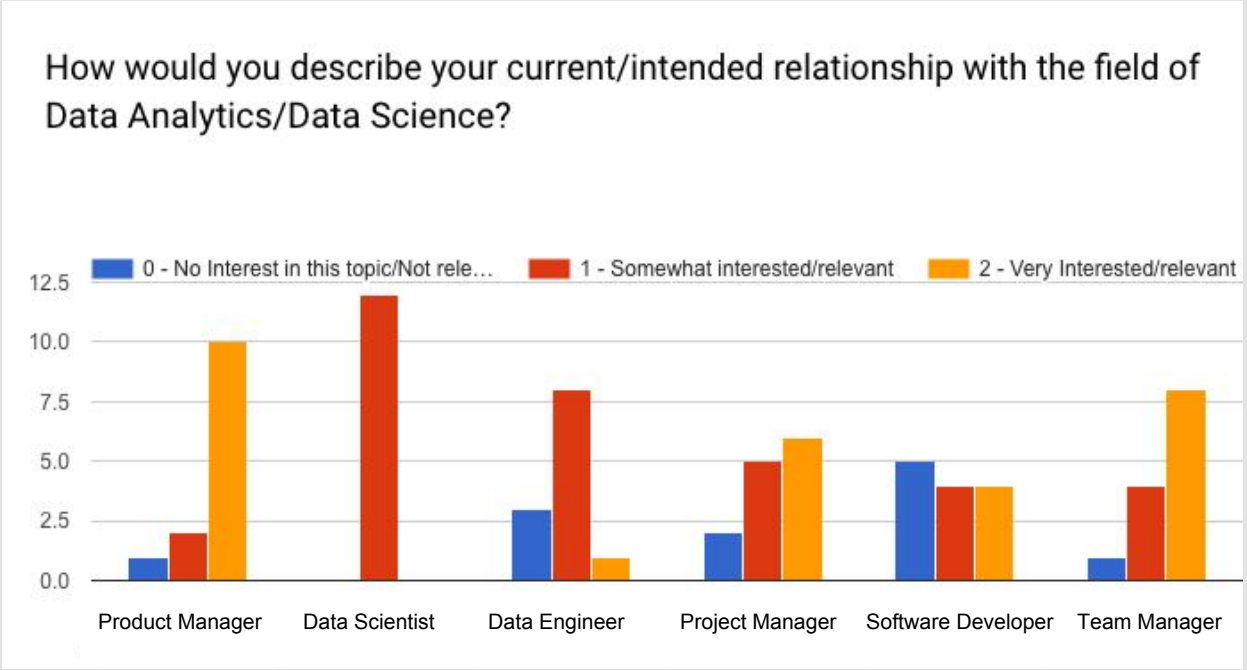
Name:	Range:
A	100 % to 94.0%
A-	< 94.0 % to 90.0%
B+	< 90.0 % to 87.0%
B	< 87.0 % to 84.0%
B-	< 84.0 % to 80.0%
C+	< 80.0 % to 77.0%
C	< 77.0 % to 74.0%
C-	< 74.0 % to 70.0%
D+	< 70.0 % to 67.0%
D	< 67.0 % to 64.0%
D-	< 64.0 % to 61.0%
R	< 61.0 % to 0.0%

Group	Weight
Homework	25%
Quizzes	20%
Class Participation	10%
Project - Individual	25%
Project - Team	20%
Total	**100%**

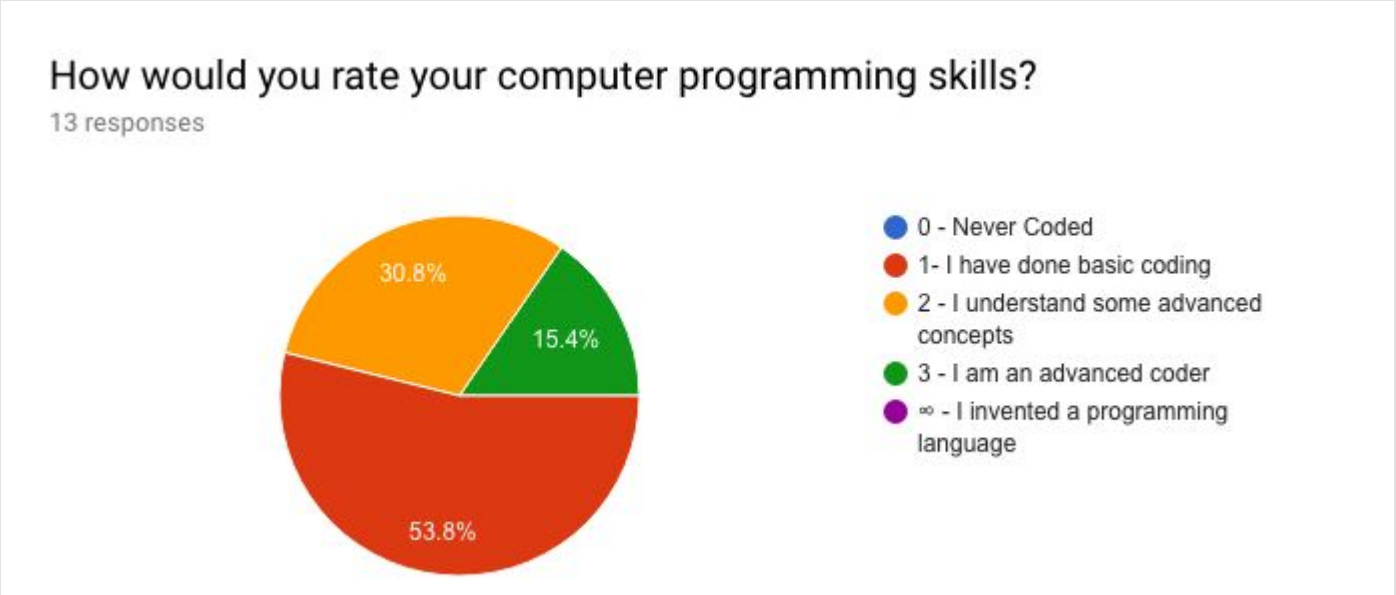
Survey Outcomes



Survey Outcomes



Survey Outcomes



Data products

49-781 Data Analytics for Product Managers

Spring 2018

What are data products?

What are Data Products?

Data Products facilitate an end-goal through the use of data.

Other products may present data to users (For example, reviewing your past order history on an Ecommerce site)

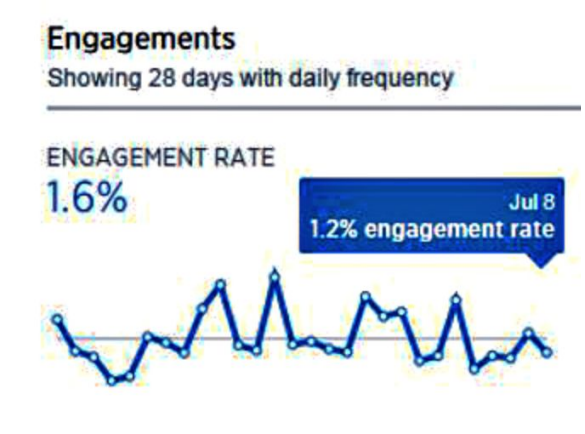
Data Products leverage data to provide actionable or useful functionality. (Showing you some other items that you may be interested in purchasing while you're reviewing order history)

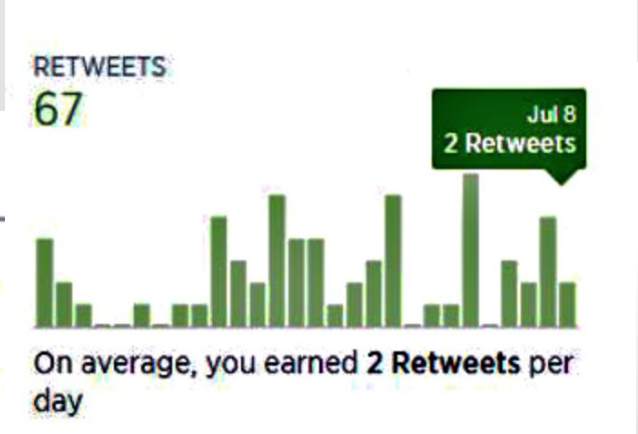
Data Analytics is a useful mechanism to enable Data Products.

Analytics is not new

The idea of analytics has been around for a long time. It was also called **Business Intelligence**

In most cases, they simply presented **summarizations** and **trends**.





What does analytics enable?

Conclusions

- You might find an occasional *conclusion* to describe the historical trends and *correlations*.
  - In the years that you spent more money on marketing, you had more sales revenue.
- But these are not intended as *predictors* of future.
  - So, what should I do?
    Should we spend more money on marketing? and how much?)
  - And, can you do that for me?
    Can the you talk directly to the budgeting software, allocate more money for marketing, see the outcomes as quickly as possible, and arrive at the most optimal proven allocation -- while go out golfing with my fellow execs?)

The New Analytics

The new analytics was born out of the idea of making *recommendations*.

It has now also extended to *implementing* recommendations, *learning* from the *outcomes* and *optimizing* the models to deliver the biggest *ROI* possible.

Can the you talk directly to the budgeting software, allocate more money for marketing, see the outcomes as quickly as possible, and arrive at the most optimal proven allocation?

More Real Time

Analytics was an *offline* or *batch* process and was rarely incorporated into real time decision making. The turnaround time was very long - primarily limited by the available technologies.

The advent of *Big Data* has allowed for rapid processing of data in *parallel*, and the ability to process streaming data in real time. This made the possibility of instantaneous meaningful recommendations real!

Benefits to Companies

Companies can directly sell these outcomes to their customers and charge them for it. For many enterprise customers, this is a great addition to their decision making process.

Companies may use this data to cause indirect outcomes. By making recommendations or placing relevant advertising, you may increase the change of a user making a purchase.

Data Products

They don't just provide insights (which is good), but actually make recommendations (which would be great!)

Their interaction with users is not just viewing, slicing and dicing - but actually guide them to a specific decision - making it a simple click or don't click decision.

They save the user from having process large amount of information and do down a decision tree. There's a unique opportunity to build trust/loyalty

Challenges with Data Products

Data Quality

Data Quality

Data Quality

Product Validation

Product Validation

Technical Infrastructure

Data Quality

Traditional analytics generally involved structured data but still had to deal with data quality issues.

Different order management systems had different required fields or different names for the same concept, etc.

Data products use highly unstructured data

Data quality is now an exponential problem

Understanding of data includes not just learning the meaning of data elements, but also data quality issues that come up as you bring large data, or data from multiple sources or domains together.

Requirements

The product has to work in the context of the user

Use has to feel comfortable about the way you are using the data

Users should feel that they are in control of the decision and the data

There's a *softer side* to data products that makes them uniquely different from traditional data analytics systems.

Data Product Iterations

Build your feature on understanding of data, data quality challenges, value proposition and ability to validate

Build your pipeline to scale the feature (which will being additional data quality challenges, value proposition and ability to validate)

Machine learning can enhance, optimize and automate the entire process

Example: LinkedIn

LinkedIn operates the world's largest professional network on the Internet with more than 467 million members in over 200 countries and territories.

They defined data product at the intersection of Big Data, User Interaction and Machine Learning

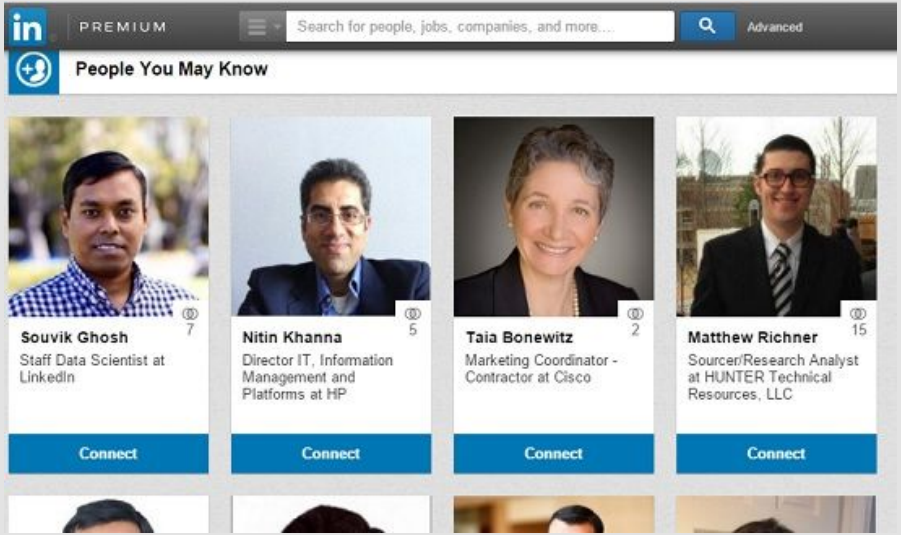
Examples

People You May Know
Skills and Endorsement
Who's viewed your profile
Groups you may like

People You may Know

PMYK gets a lot of attention as a data product

Let's explore how it might work…



The Feature

Describe the capability/feature that the data product is enabling for the company and how the company benefits from that feature.

The Data

Describe the data used to generate that feature, including how the data is connected together to evolve the feature.

The Challenges

Describe the type of challenges the company may have gone through in the process - dealing with data quality, consistency etc.

The Outcome

Describe the outcome of the feature and how is it presented to customers/users.

The Validation

Describe how they possibly test that the feature is working as expected, and what data they may use to analyze the performance of the feature and make improvements

Elementary Statistics

49-781 Data Analytics for Product Managers

Spring 2018

Elementary Statistics

In statistics, we turn data into information. We can use it to identify trends and understand features of populations often by analyzing a sample.

Variables

Variables are used to describe the characteristics of an individual in a population.

The idea is that the value of this variable or variables changes between individuals in the population giving them some uniqueness of behavior.

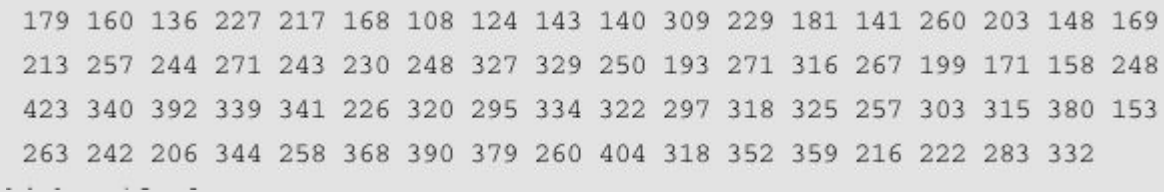
For example, below are the girth, height and volume of some felled black cherry trees.

```
##   Girth Height Volume
## 1   8.3     70   10.3
## 2   8.6     65   10.3
## 3   8.8     63   10.2
## 4  10.5     72   16.4
## 5  10.7     81   18.8
```

Variables

Variables can be numeric or categorical. A numeric variable takes number values and can be continuous or discrete. A continuous numeric variable can take any real value in a range

- For example, the weight of different chicks in a farm



- A discrete numeric variable takes distinct numerical values.
- For example, the Run number in the above experiment

Variables

Categorical variables take only finite number of possibilities and they are not numeric.
Categorical variables that can be ranked are called ordinal variables
Those that cannot be ranked or have no order are called nominal variables.
Example of a nominal variable is chick feed

```
 [1] horsebean horsebean horsebean horsebean horsebean horsebean horsebean
 [8] horsebean horsebean horsebean linseed   linseed   linseed   linseed
[15] linseed   linseed   linseed   linseed   linseed   linseed   linseed
[22] linseed   soybean   soybean   soybean   soybean   soybean   soybean
[29] soybean   soybean   soybean   soybean   soybean   soybean   soybean
[36] soybean   sunflower sunflower sunflower sunflower sunflower sunflower
[43] sunflower sunflower sunflower sunflower sunflower sunflower meatmeal
[50] meatmeal  meatmeal  meatmeal  meatmeal  meatmeal  meatmeal  meatmeal
[57] meatmeal  meatmeal  meatmeal  casein    casein    casein    casein
[64] casein    casein    casein    casein    casein    casein    casein
```

Example of an ordinal variable is "dose of a drug" as Low, Medium and High

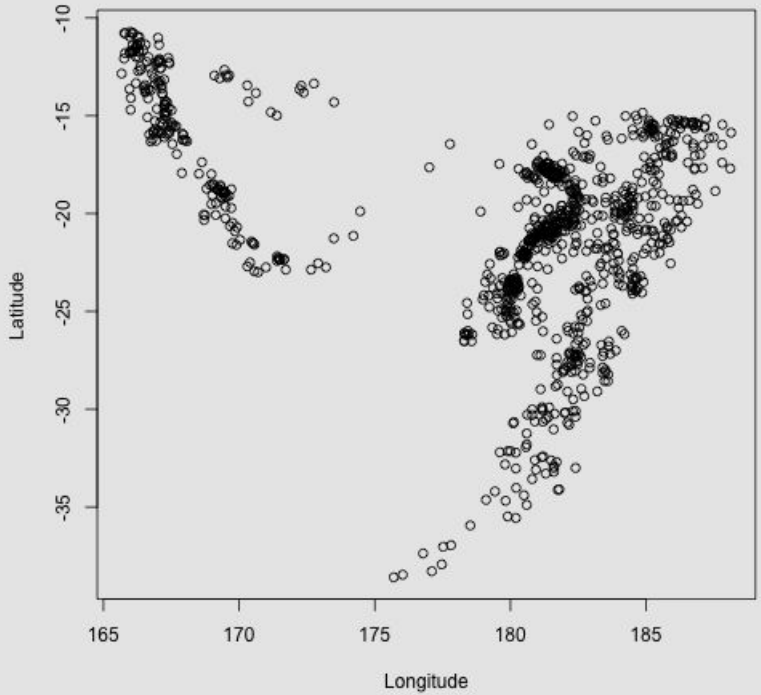
Univariate and Multivariate Data

When data is in a single dimension, it has a single variable and is referred to as univariate data.
When data represents multiple variables it's called multivariate data.
You use multivariate data when a single dimension does not sufficietnly express the individuals in the population

For example, Latitude and Longitude together represent a location

lat	long	c
-20.42	181.62	
-20.62	181.03	
-26.00	184.10	
-17.97	181.66	
-20.42	181.96	

Plotting Locations of Earthquakes



Parameters and Statistics

The characteristics of an entire population are known as parameters.We typically collect data about a sample of the population.Estimating the parameters of interest for this sample are known as statistics.

Example: Distribution of Hair and Eye color across the United States

We have a sample of smaller set of people from which we can calculate statistics and draw conclusion on the overall population

```
##        Eye
## Hair    Brown Blue Hazel Green
##   Black    32   11    10     3
##   Brown    53   50    25    15
##   Red      10   10     7     7
##   Blond     3   30     5     8
```

Exercise

For each of the following, identify the type of variable described: numeric-continuous, numeric-discrete, categorical-nominal, or categorical-ordinal:

- The number of blemishes on the hood of a car coming off a production line
- A survey question that asks the participant to select from Strongly agree, Agree, Neutral, Disagree, and Strongly disagree
- The noise level (in decibels) at a concert
- The noise level out of three possible choices: high, medium, low
- A choice of primary color

Exercise with Answers

For each of the following, identify the type of variable described: numeric-continuous, numeric-discrete, categorical-nominal, or categorical-ordinal:

- The number of blemishes on the hood of a car coming off a production line *(Numeric-discrete)*
- A survey question that asks the participant to select from Strongly agree, Agree, Neutral, Disagree, and Strongly disagree *(Categorical-ordinal)*
- The noise level (in decibels) at a concert *(Numeric-continuous)*
- The noise level out of three possible choices: high, medium, low *(Categorical-ordinal)*
- A choice of primary color *(Categorical-nominal)*

Exercise

For each of the following, identify whether the quantity discussed is a population parameter or a sample statistic. If the latter, also identify what the corresponding population parameter is.

- The percentage of 50 New Zealanders who own a gaming console
- The average number of blemishes found on the hoods of three cars in the No Dodgy Carz yard
- The proportion of domestic cats in the United States that wear a collar
- The average number of times per day a vending machine is used in a year
- The average number of times per day a vending machine is used in a year, based on data collected on three distinct days in that year

Exercise with Answers

For each of the following, identify whether the quantity discussed is a population parameter or a sample statistic. If the latter, also identify what the corresponding population parameter is.

- The percentage of 50 New Zealanders who own a gaming console *(Sample statistic. Population parameter is the proportion of NZers who own a gaming console.)*
- The average number of blemishes found on the hoods of three cars in the No Dodgy Carz yard *(Sample statistic. Population parameter is the average number of blemishes on the hoods of all cars at No Dodgy Carz.)*
- The proportion of domestic cats in the United States that wear a collar *(Population parameter)*
- The average number of times per day a vending machine is used in a year *(Population parameter)*
- The average number of times per day a vending machine is used in a year, based on data collected on three distinct days in that year *(Sample statistic. Population parameter is the previous question)*

Summary Statistics

These types of statistics are used to summarize the variable of a sample - and attempt to simplify them to a single or a small set of numbers.

“Measures of Centrality” - Focus on where numeric observations are centered

Arithmetic Mean

Arithmetic Mean: A “balance point” of a collection observations. Add up all the observations and divide by the number of observations.

$$\bar{x} = \frac{(x_1 + x_2 + \ldots + x_n)}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\frac{2 + 4.4 + 3 + 3 + 2 + 2.2 + 2 + 4}{8} = 2.825$$

Median

A “middle magnitude” of your observations. Place your observations in order and find the middle observation (If even number, find the mean of the two middle values)

$$\bar{m}_x = \begin{cases} x_i^{\left(\frac{n+1}{2}\right)}, & \text{if } n \text{ is odd} \\ \left(x_i^{\left(\frac{n}{2}\right)} + x_j^{\left(\frac{n}{2}+1\right)}\right)/2, & \text{if } n \text{ is even} \end{cases}$$

2, 2, 2, 2.2, 3, 3, 4, 4.4

$$\left(x_i^{(4)} + x_j^{(5)}\right)/2 = (2.2 + 3)/2 = 2.6$$

Mode

The “most common” observation. You can have data with no mode (all unique) or multiple modes (more than one value occurs most times) This is used with numeric discrete data

2, 2, 2, 2.2, 3, 3, 4, 4.4

Observation	2	2.2	3	4	4.4
Frequency	3	1	2	1	1

Here, 2 is the value and 3 is the frequency of that value.

Counts & Proportions

Counts typically works for categorical variables and counts the number of times a particular value occurs

Proportion (calculated assuming the total is 1) is the portion of a observations that have a value or that fall within a range.

(Percentage is an alternate for proportion - calculated out of 100)

Look at some data

Chicken weight by feed type

Count

```
##
##    casein horsebean   linseed  meatmeal   soybean sunflower
##        12        10        12        11        14        12
```

Proportions

```
##
##    casein horsebean   linseed  meatmeal   soybean sunflower
## 0.1690141 0.1408451 0.1690141 0.1549296 0.1971831 0.1690141
```

Percentages

```
##
##    casein horsebean   linseed  meatmeal   soybean sunflower
##  16.90141  14.08451  16.90141  15.49296  19.71831  16.90141
```

weight	feed
179	horsebean
160	horsebean
136	horsebean
227	horsebean
217	horsebean
168	horsebean
108	horsebean
124	horsebean
143	horsebean
140	horsebean
309	linseed
229	linseed
181	linseed
141	linseed
260	linseed
203	linseed
148	linseed
169	linseed
213	linseed
257	linseed
244	linseed
271	linseed
243	soybean

Quantiles

A quantile indicates an observation's rank when compared to all other available observations.

A median is an example of a quantile where half the measurements like below it (and half above it) This is the 0.5th quantile (Quantiles can also be expressed as percentile - so 0.5th quantile would be 50th percentile)

80% quantile for chicken weights 332

Five Number Summary

Five number summary comprises of the 0th percentile (the minimum), the 25th percentile, the 50th percentile, the 75th percentile, and the 100th percentile (the maximum).

The 0.25th quantile is referred to as the first or lower quartile, and the 0.75th quantile is referred to as the third or upper quartile.

The median is the second quartile, with the maximum value being the fourth quartile.

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 108.0   204.5   258.0   261.3   323.5   423.0 
```

Interpreting Five Number Summaries

Chicken weight data

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   108.0   204.5   258.0   261.3   323.5   423.0
```

The difference between median and minimum is comparable to the median and the maximum - indicating an even spread across the range

Earthquake data

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00    4.30    4.60    4.62    4.90    6.40
```

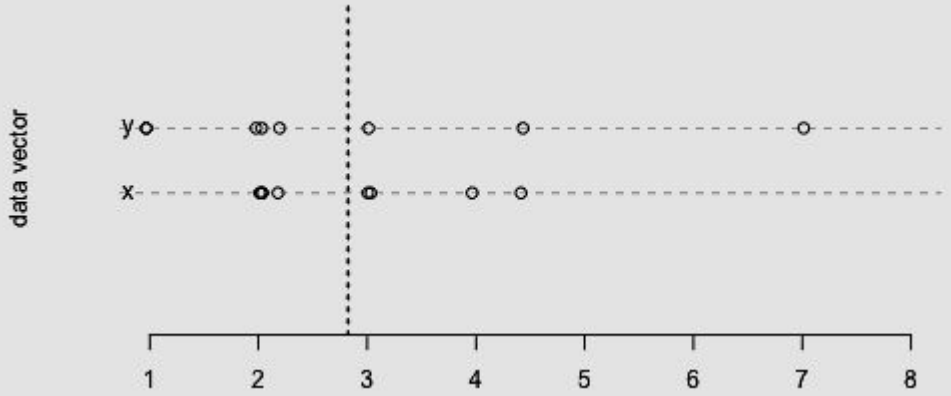
The difference between median and minimum is much smaller than the median and the maximum - indicating larger spread at higher end of the earthquakes.

More Spread

x = 2, 4.4, 3, 3, 2, 2.2, 2, 4

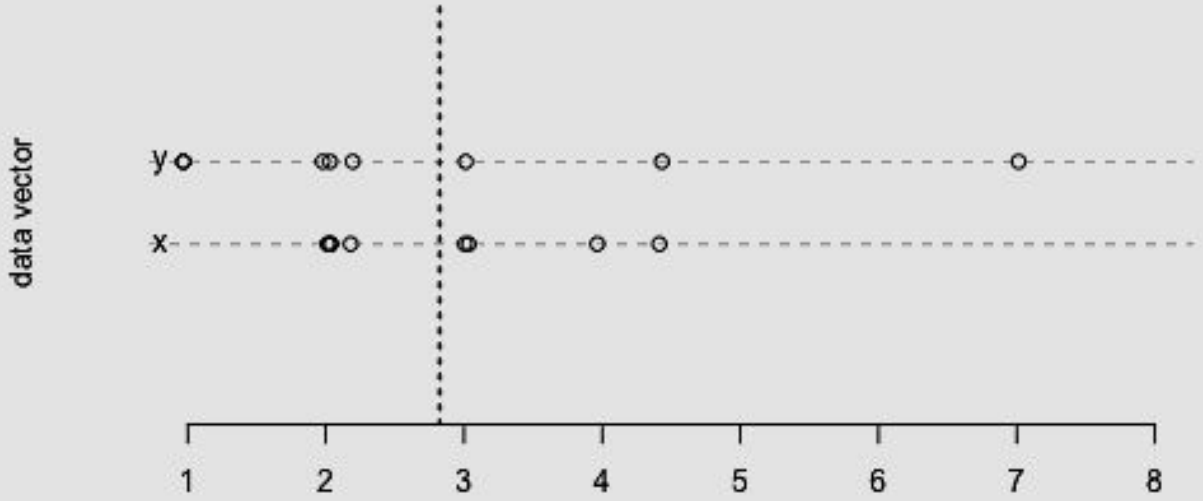
y = 1, 4.4, 1, 3, 2, 2.2, 2, 7

mean of x is 2.825 and mean of y is 2.825 which are the same, but are the data similar?



Let's quantify the spread

Though the means are same, the observations of y are more spread out and the measures of centrality don't reflect this spread.



Quantifying spread

You can use the sample variance to measure measures the degree of the spread of observations around their arithmetic mean. It is a representation of the average squared distance of each observation when compared to the mean.

$$s_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$s_x = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Standard Deviation is the square root of Variance:

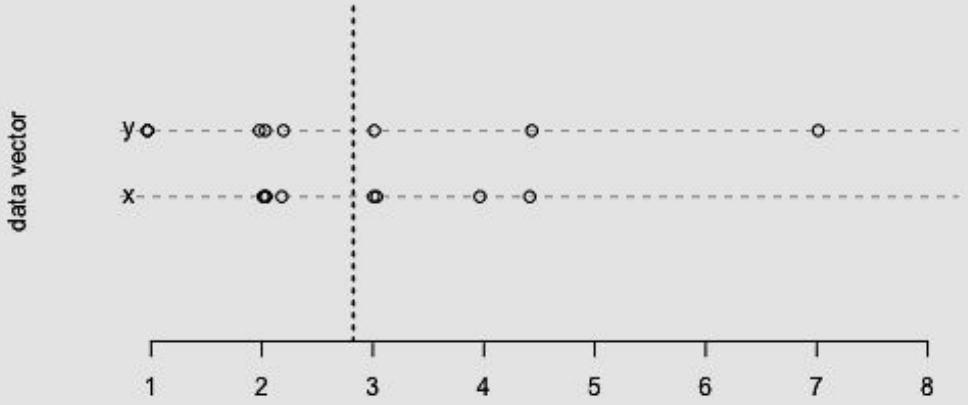
Another measure of spread

Interquartile range is measured by the width of the middle 50% of the data - the difference between upper (75%) and lower (25%) quartile.

Comparing spreads

For x (2, 4.4, 3, 3, 2, 2.2, 2, 4), the variance is 0.9078571, standard deviation is 0.9528154 and IQR is 1.25

For y (1, 4.4, 1, 3, 2, 2.2, 2, 7), the variance is 4.0507143, standard deviation is 2.0126386 and IQR is 1.6



Interpreting Standard Deviation

For the chicken weights, standard deviation is 78.0736999 which indicates that on 'average', the weight of a chicken is around 78.0736999 grams away from the mean weight, which is 261.3098592

Relationships between Variables

Here's data about age and circumference of orange trees

```
##     age circumference
## 1   118            30
## 2   484            58
## 3   664            87
## 4  1004           115
## 5  1231           120
## 6  1372           142
## 7  1582           145
## 8   118            33
## 9   484            69
## 10  664           111
## 11 1004           156
## 12 1231           172
## 13 1372           203
## 14 1582           203
## 15  118            30
## 16  484            51
## 17  664            75
## 18 1004           108
## 19 1231           115
## 20 1372           139
## 21 1582           140
## 22  118            32
```

Covariance

Covariance describes how much two variables change "together" and whether they have a similar or opposite effect on each other.

$$r_{xy} = \frac{1}{n-1} \sum_{i-1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

x = 2, 4.4, 3, 3, 2, 2.2, 2, 4

y = 1, 4.4, 1, 3, 2, 2.2, 2, 7

Covariance = 1.4792857

A positive value indicates that the two variables change together in the same direction. A negative value indicates that they change in opposite directions. A zero value indicates no linear relationship.

Correlation

Correlation allows you to interpret the covariance further by identifying both the direction and strength of any association.

Pearson's Correlation Coefficient: Divide the sample covariance by the product of the standard deviation of each dat set.

$$\rho_{xy} = \frac{r_{xy}}{s_x s_y}$$

This ensures that $-1 \leq \rho_{xy} \leq 1$

When it is -1, there exists a perfect negative linear relationship and when it is 1, a perfect positive linear relationship exists.

Correlation Coefficient for our Data Sets

2, 4.4, 3, 3, 2, 2.2, 2, 4 and 1, 4.4, 1, 3, 2, 2.2, 2, 7

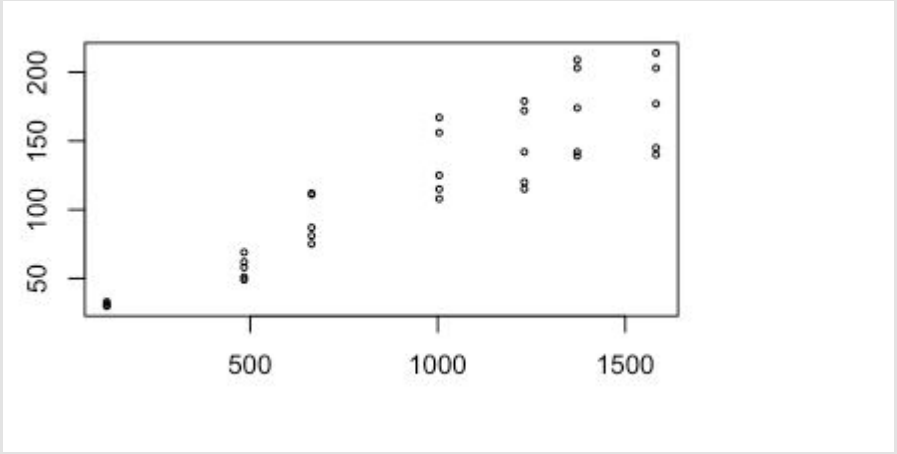
Covariance is 1.4792857

Standard Deviations are 0.9528154 and 2.0126386

Correlation Coefficient is 0.7713962, which indicates a moderate to strong positive relationship.

Orange Trees

```
##    age circumference
## 1  118            30
## 2  484            58
## 3  664            87
## 4 1004           115
## 5 1231           120
```

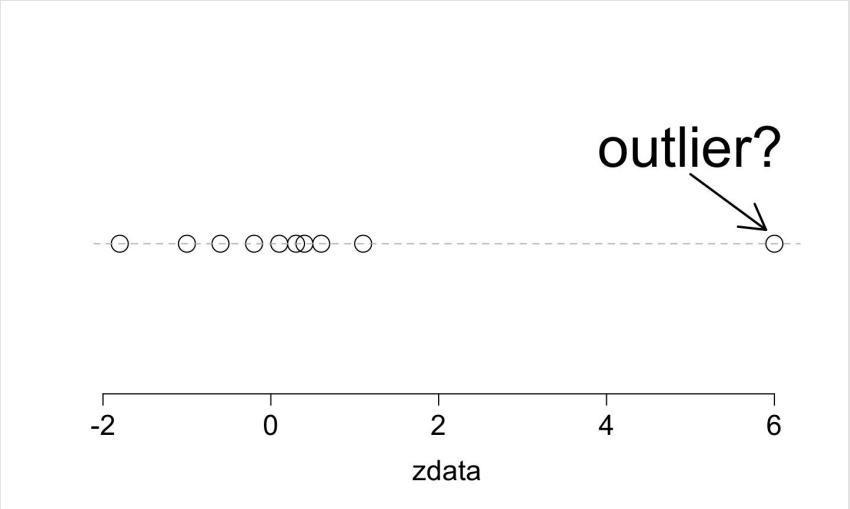


Correlation is 0.9135189 which shows a strong positive correlation

Outliers

An observation that does not appear to "fit" with the rest of the data.

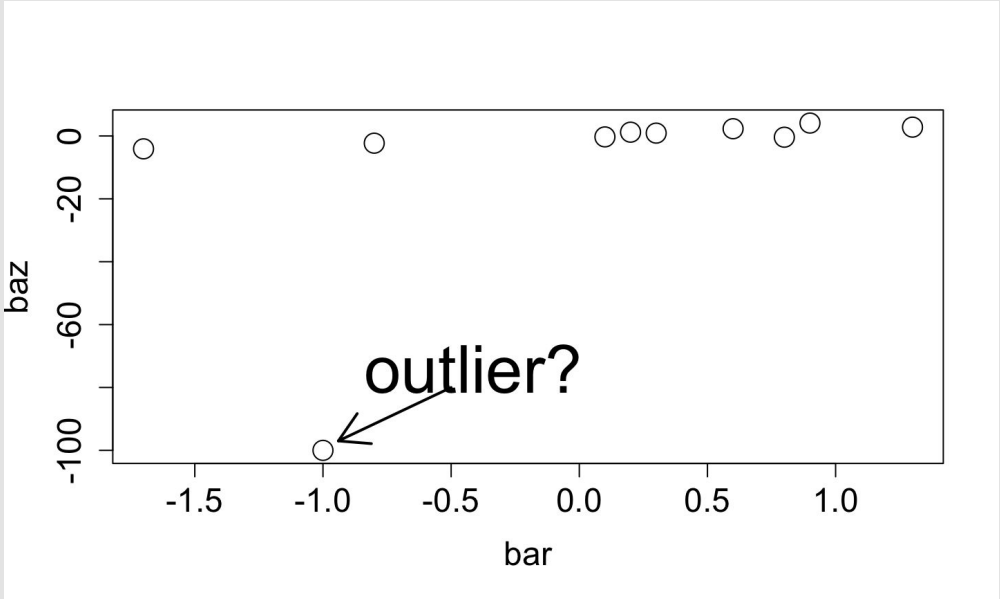
0.6, -0.6, 0.1, -0.2, -1, 0.4, 0.3, -1.8, 1.1, 6



Outliers in bivariate data

0.1, 0.3, 1.3, 0.6, 0.2, -1.7, 0.8, 0.9, -0.8, -1

-0.3, 0.9, 2.8, 2.3, 1.2, -4.1, -0.4, 4.1, -2.3, -100



Dealing with Outliers

Outliers may occur naturally in the population. But they may also occur due to contamination of the sample.

It is a common practice is omit outliers for data analysis, but it's not always easy in practice without knowing the cause of the outliers.

Impact of outliers:

0.6, -0.6, 0.1, -0.2, -1, 0.4, 0.3, -1.8, 1.1, 6

Mean with all data: 0.49
Mean with outlier removed: -0.1222222

The impact of a single observation is significant.

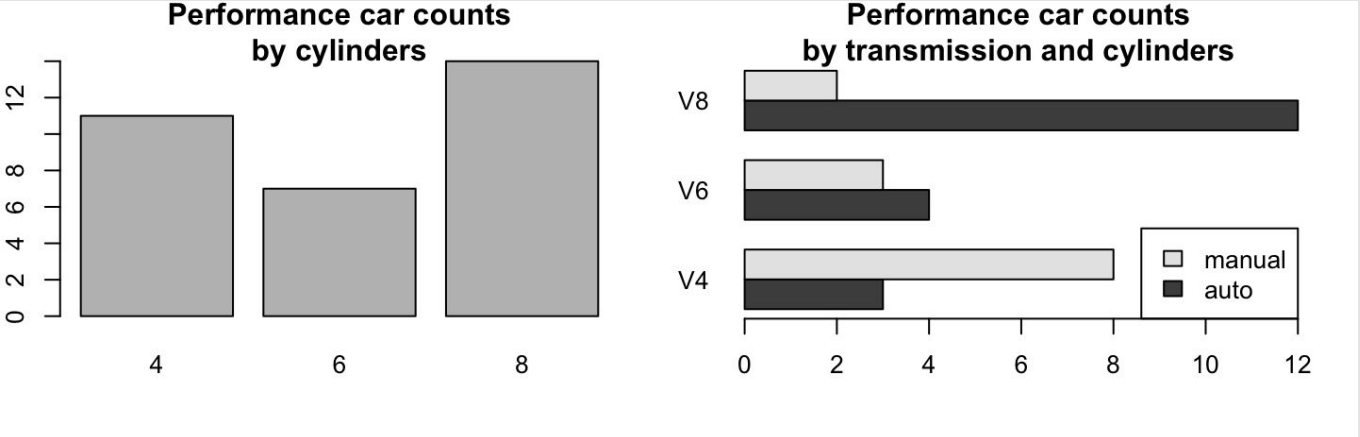
Elementary Data Visualization

49-781 Data Analytics for Product Managers

Spring 2018

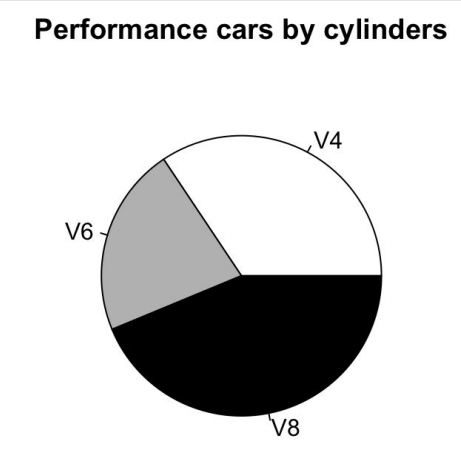
Barplots

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
```



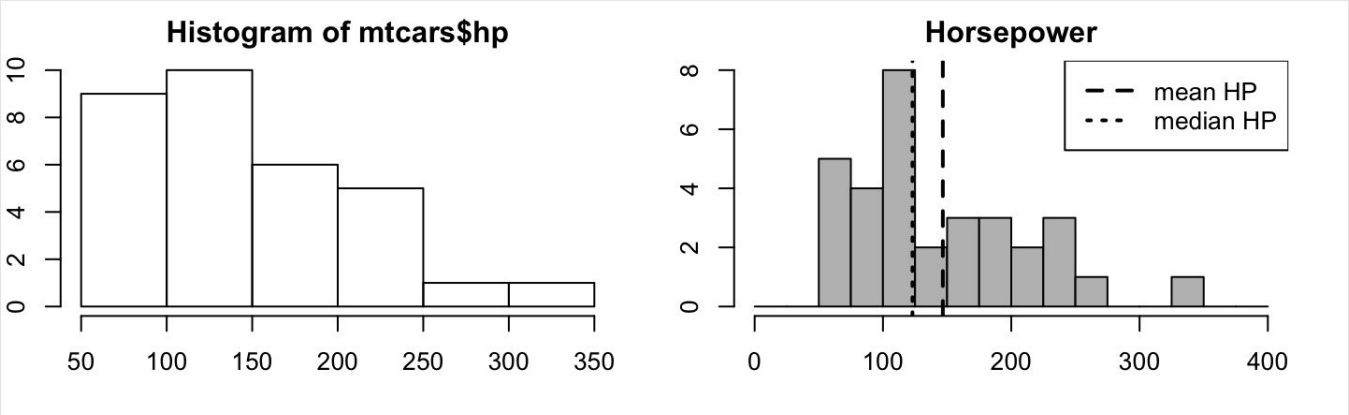
Pie Charts

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
```

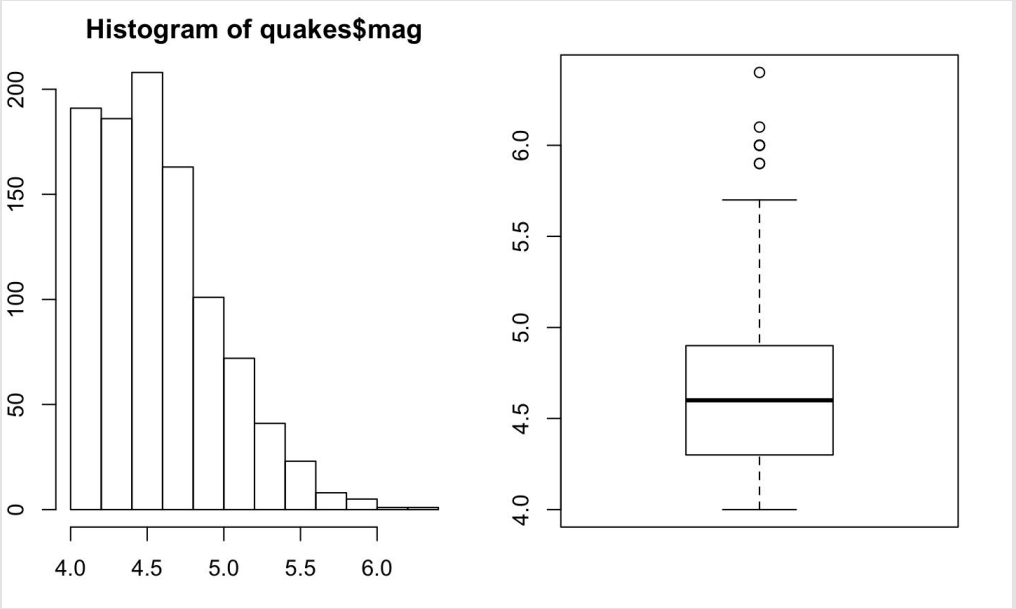


Histograms

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
```



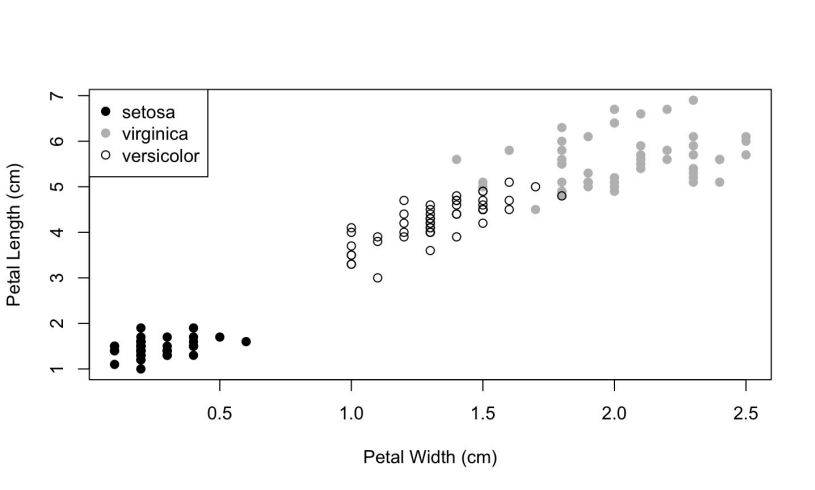
Boxplots



By default, boxplot defines an outlier as an observation that lies more than 1.5 times the IQR below the lower quartile or above the upper quartile.

Scatterplot

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
```



Lab Exercise

49-781 Data Analytics for Product Managers
Spring 2018

Development Environment

- You can use a stand-alone IDE like PyCharm or use Jupyter notebook for your development
  - However, all your code submissions should be stand-alone files that I can run with PyCharm
  - Pycharm can automatically download most libraries as you use them so it should be just as easy as jupyter
- Install PyCharm now
- Download CSV files used in the following slides from Canvas files
- Start coding…

Scatterplot

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt


###################
q1=pd.read_csv("quakes.csv")
print(q1)
plt.scatter(q1["lat"], q1["long"], alpha=0.5)
plt.show()
###################
```

Central Tendencies

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt


##########
d = pd.Series([5,7,2,4,5,6,4,5,6,4,3,5,6,5,3])
print("Mean: ",d.mean())
print("Mode: ", d.mode())
print("Median: ", d.median())
#########
```

Spreads

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt


###########
c1=pd.read_csv("chickwts.csv")
w1 = c1["weight"]
print(w1.mean())
print(w1.median())
f1 = c1["feed"]
print(f1.value_counts())
print(f1.value_counts()/f1.size)
print(f1.value_counts()/f1.size*100)
print(w1.quantile(0.8))
print(w1.describe())
#######
```

Spreads

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt


#######
q1=pd.read_csv("quakes.csv")
print(q1["mag"].describe())
print(q1.describe())
########
```

Correlation

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt


######
o = pd.read_csv("orange.csv")
o1 = o[["age","circumference"]]
print(o1)
print("Correlation: ",o1.corr())
plt.scatter(o1["age"],o1["circumference"], alpha=0.5)
plt.show()
########
```

Barplot

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt


########
m1=pd.read_csv("mtcars.csv")
cf = m1["cyl"].value_counts()
cd=cf.apply(pd.Series)
cd.columns=["freq"]
print(cd)
plt.bar(cd.index,cd["freq"])
plt.show()
############
```

Pie chart

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt


########
m1=pd.read_csv("mtcars.csv")
cf = m1["cyl"].value_counts()
cd=cf.apply(pd.Series)
cd.columns=["freq"]
print(cd)
plt.pie(cd["freq"], labels=cd.index, startangle=90)
plt.axis('equal')
plt.show()
############
```

Boxplot

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt


######
o = pd.read_csv("orange.csv")
o1 = o[["age","circumference"]]
print(o1)
plt.boxplot(o1["age"])
plt.show()
########
```