Slide 1

Python Code

import pandas as pd

import numpy

from matplotlib import pyplot as plt

pd.options.display.max\_columns=300

x=pd.read\_excel("G:/Fortune500Companies\_Case Study.xlsx",0)

x.info()

Revenue Graph

x=plt.hist(x['Revenue ($ Millions)'], color = 'orange', edgecolor = 'black',

bins = int(180/5))

plt.gcf().subplots\_adjust(bottom=0.5) # ensures enough space at bottom of plot for text

plt.title('Revenue ($ Millions)', fontsize=18)

plt.ylabel('# of Companies')

plt.xlabel('Revenue, in Millions')

plt.xticks(rotation='vertical')

plt.show()

Assets Graph

x=plt.hist(x['Total Assets ($ Millions)'], color = 'orange', edgecolor = 'black',

bins = int(180/5))

plt.gcf().subplots\_adjust(bottom=0.5) # ensures enough space at bottom of plot for text

plt.title('Total Assets ($ Millions)', fontsize=18)

plt.ylabel('# of Companies')

plt.xlabel('Total Assets, in Millions')

plt.xticks(rotation='vertical')

plt.show()

Calculations

TotalRevenue = x['Revenue ($ Millions)'].sum()

print(TotalRevenue)

MedianRevenue = x['Revenue ($ Millions)'].median()

print(MedianRevenue)

TotalAssets = x['Total Assets ($ Millions)'].sum()

print(TotalAssets)

MedianAssets = x['Total Assets ($ Millions)'].median()

print(MedianAssets)

Slide 2

GICS Sector Graph – Python Code

x=plt.hist(x['GICS Sector'], color = 'blue', edgecolor = 'black',

bins = int(180/5))

plt.gcf().subplots\_adjust(bottom=0.5) # ensures enough space at bottom of plot for text

plt.title('GICS Sector', fontsize=18)

plt.ylabel('Count')

plt.xticks(rotation='vertical')

plt.show()

Slide 3

# SQL code below creates SUM of Revenue by Company on Sheet 3, and EOP data on Sheet 2, with Sheet 1 to give revenue earned by company and EOP data against other variables on Sheet 1

SQL Code

SELECT Sum(`'WF Revenue$'`.Revenue), `'WF Revenue$'`.`Company Name`, `'WF Revenue$'`.`Company ID`, `F500$`.`EPS % Change (10 year)`, `F500$`.`EPS % Change (5 year)`, `F500$`.`EPS % Change (from 2012)`, `F500$`.`Market Value on March 31, 2014 ($ Millions)`, `F500$`.`Profit ($ Millions)`, `F500$`.`Profit (% change from previous year)`, `F500$`.`Profit as a % of Assets`, `F500$`.`Profit as a % of Sales`, `F500$`.`Profit as a % of Stockholders' Equity`, `F500$`.`Revenue ($ Millions)`, `F500$`.`Revenue (% change from previous year)`, `F500$`.`Total Assets ($ Millions)`, `F500$`.`Total Return to Investors`, `F500$`.`Total Return to Investors (10 year, annualized)`, `F500$`.`Total Return to Investors (5 year, annualized)`, `F500$`.`Total Shareholder Equity ($ Millions)`, `'WF Exposure$'`.`EOP Commitment`, `'WF Exposure$'`.`EOP Outstanding`

FROM `C:\Users\schemsak\Desktop\Fortune500Companies\_Case Study.xlsx`.`F500$` `F500$`, `C:\Users\schemsak\Desktop\Fortune500Companies\_Case Study.xlsx`.`'WF Exposure$'` `'WF Exposure$'`, `C:\Users\schemsak\Desktop\Fortune500Companies\_Case Study.xlsx`.`'WF Revenue$'` `'WF Revenue$'`

WHERE `F500$`.`Company ID` = `'WF Revenue$'`.`Company ID` AND `'WF Exposure$'`.`Legal Entity ID` = `F500$`.`Company ID`

GROUP BY `'WF Revenue$'`.`Company Name`, `'WF Revenue$'`.`Company ID`, `F500$`.`EPS % Change (10 year)`, `F500$`.`EPS % Change (5 year)`, `F500$`.`EPS % Change (from 2012)`, `F500$`.`Market Value on March 31, 2014 ($ Millions)`, `F500$`.`Profit ($ Millions)`, `F500$`.`Profit (% change from previous year)`, `F500$`.`Profit as a % of Assets`, `F500$`.`Profit as a % of Sales`, `F500$`.`Profit as a % of Stockholders' Equity`, `F500$`.`Revenue ($ Millions)`, `F500$`.`Revenue (% change from previous year)`, `F500$`.`Total Assets ($ Millions)`, `F500$`.`Total Return to Investors`, `F500$`.`Total Return to Investors (10 year, annualized)`, `F500$`.`Total Return to Investors (5 year, annualized)`, `F500$`.`Total Shareholder Equity ($ Millions)`, `'WF Exposure$'`.`EOP Commitment`, `'WF Exposure$'`.`EOP Outstanding`

ORDER BY Sum(`'WF Revenue$'`.Revenue) DESC

Slide 4

R Code for 3D Pie Chart

# Query added into new Excel tab. Top 5 taken from Revenue above.

Library(readxl)

mydata3 <- read\_excel("C:/Users/chemsak/desktop/Wells Fargo/Fortune500Companies\_Case Study\_mod.xlsx",5)

library(plotrix)

slices <- c(63604899.41, 44865163.144, 16569328.831, 16149895.612, 15962458.173)

lbls <- c("Regions Financial Corp., 63.5M", "Vulcan Materials, 44.9M", "Murphy Oil, 16.6M", "Microchip Technology, 16.1M", "Wal-Mart Stores, 16M")

pie3D(slices,labels=lbls,explode=0.1,

main="Top 5 Companies for WF Revenue ")

Slide 5,6,7

R Code

Sources of Revenue Bar Graph

**library(rpivotTable)**

**library(readxl)**

**mydata2 <- read\_excel("G:/Fortune500Companies\_Case Study.xlsx",2)**

**rpivotTable(mydata2)**

Slide 7

mydata3 <- read\_excel("G:/Copy of Fortune500Companies\_Case Study\_Numbers\_2.xlsx",4) # sheet # from SQL query merging all data including EOP data.

newdata <- mydata3[c(2,7,8,9,14,20:22)]

P<-cor(newdata,use="complete.obs")

head(round(P,2))

corrplot(P, method="circle")

Regression Model: Profit, Market Value, Shareholder Equity, and Revenue

> model<-lm(newdata$'Sum(Revenue)'~newdata$'Profit ($ Millions)'+newdata$'Market Value on March 31, 2014 ($ Millions)'+newdata$'Total Shareholder Equity ($ Millions)'+newdata$'Revenue ($ Millions)', data=newdata)

> summary(model)

Call:

lm(formula = newdata$"Sum(Revenue)" ~ newdata$"Profit ($ Millions)" +

newdata$"Market Value on March 31, 2014 ($ Millions)" + newdata$"Total Shareholder Equity ($ Millions)" +

newdata$"Revenue ($ Millions)", data = newdata)

Residuals:

Min 1Q Median 3Q Max

-8735827 -483411 -87967 707793 13428219

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 126161.794 86570.310 1.457 0.145699

newdata$"Profit ($ Millions)" -53.518 16.018 -3.341 0.000901 \*\*\*

newdata$"Market Value on March 31, 2014 ($ Millions)" -4.275 2.209 -1.935 0.053551 .

newdata$"Total Shareholder Equity ($ Millions)" -1.196 4.204 -0.284 0.776169

newdata$"Revenue ($ Millions)" 109.290 2.499 43.741 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1564000 on 465 degrees of freedom

(31 observations deleted due to missingness)

Multiple R-squared: 0.875, Adjusted R-squared: 0.8739

F-statistic: 813.4 on 4 and 465 DF, p-value: < 2.2e-16

Regression Model: Profit and Revenue

> model<-lm(newdata$'Sum(Revenue)'~newdata$'Profit ($ Millions)'+newdata$'Revenue ($ Millions)', data=newdata)

> summary(model)

Call:

lm(formula = newdata$"Sum(Revenue)" ~ newdata$"Profit ($ Millions)" +

newdata$"Revenue ($ Millions)", data = newdata)

Residuals:

Min 1Q Median 3Q Max

-8801989 -462943 -115399 672809 14065669

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 92802.218 82285.585 1.128 0.26

newdata$"Profit ($ Millions)" -70.608 14.814 -4.766 2.47e-06 \*\*\*

newdata$"Revenue ($ Millions)" 106.190 2.143 49.545 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1567000 on 497 degrees of freedom

(1 observation deleted due to missingness)

Multiple R-squared: 0.8678, Adjusted R-squared: 0.8673

F-statistic: 1632 on 2 and 497 DF, p-value: < 2.2e-16

Regression Model: Profit and Revenue

> model<-lm(newdata$'Sum(Revenue)'~newdata$'Revenue ($ Millions)', data=newdata)

> summary(model)

Call:

lm(formula = newdata$"Sum(Revenue)" ~ newdata$"Revenue ($ Millions)",

data = newdata)

Residuals:

Min 1Q Median 3Q Max

-9024236 -456546 -109904 686927 15711065

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 82071.867 84029.309 0.977 0.329

newdata$"Revenue ($ Millions)" 100.383 1.801 55.725 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1601000 on 498 degrees of freedom

(1 observation deleted due to missingness)

Multiple R-squared: 0.8618, Adjusted R-squared: 0.8615

F-statistic: 3105 on 1 and 498 DF, p-value: < 2.2e-16

> model<-lm(newdata$'Sum(Revenue)'~newdata$'Market Value on March 31, 2014 ($ Millions)'+newdata$'Revenue ($ Millions)', data=newdata)

> summary(model)

Call:

lm(formula = newdata$"Sum(Revenue)" ~ newdata$"Market Value on March 31, 2014 ($ Millions)" +

newdata$"Revenue ($ Millions)", data = newdata)

Residuals:

Min 1Q Median 3Q Max

-8935860.5 -497279.7 -91417.1 722825.8 14292657.9

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 139460.462235 87117.018830 1.60084 0.11009

newdata$"Market Value on March 31, 2014 ($ Millions)" -6.872570 1.691417 -4.06320 0.00005676 \*\*\*

newdata$"Revenue ($ Millions)" 106.801609 2.377652 44.91894 < 0.000000000000000222 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1577791 on 468 degrees of freedom

(30 observations deleted due to missingness)

Multiple R-squared: 0.8718635, Adjusted R-squared: 0.8713159

F-statistic: 1592.178 on 2 and 468 DF, p-value: < 0.00000000000000022204

Bonus Slide – Python Code

import pandas as pd

import numpy

import pandas as pd

import numpy

# The model should automatically split on the continuous variables where it gauges performance will be best.

x=pd.read\_excel("G:/Copy of Fortune500Companies\_Case Study\_Numbers\_2.xlsx",3)

xrenamed=x.rename(index=str, columns={"Sum(Revenue)": "SumRevenue","Profit ($ Millions)": "Profit", "Revenue ($ Millions)":"Revenue" })

y=xrenamed.SumRevenue

x\_predictors=['Revenue','Profit']

xpredict=xrenamed[x\_predictors]

from sklearn.tree import DecisionTreeRegressor

x\_model = DecisionTreeRegressor()

x\_model.fit(xpredict, y)

print(x\_model.predict(xpredict.head())) #predicted values on original data

from sklearn.model\_selection import train\_test\_split

train\_X, val\_X, train\_y, val\_y = train\_test\_split(xpredict, y,random\_state = 0)

x\_model.fit(train\_X, train\_y)

from sklearn.metrics import mean\_absolute\_error

val\_predictions = x\_model.predict(val\_X)

print(mean\_absolute\_error(val\_y, val\_predictions))

[63604899.409 44865163.144 16569328.831 16149895.612 15962458.173]

870350.24936