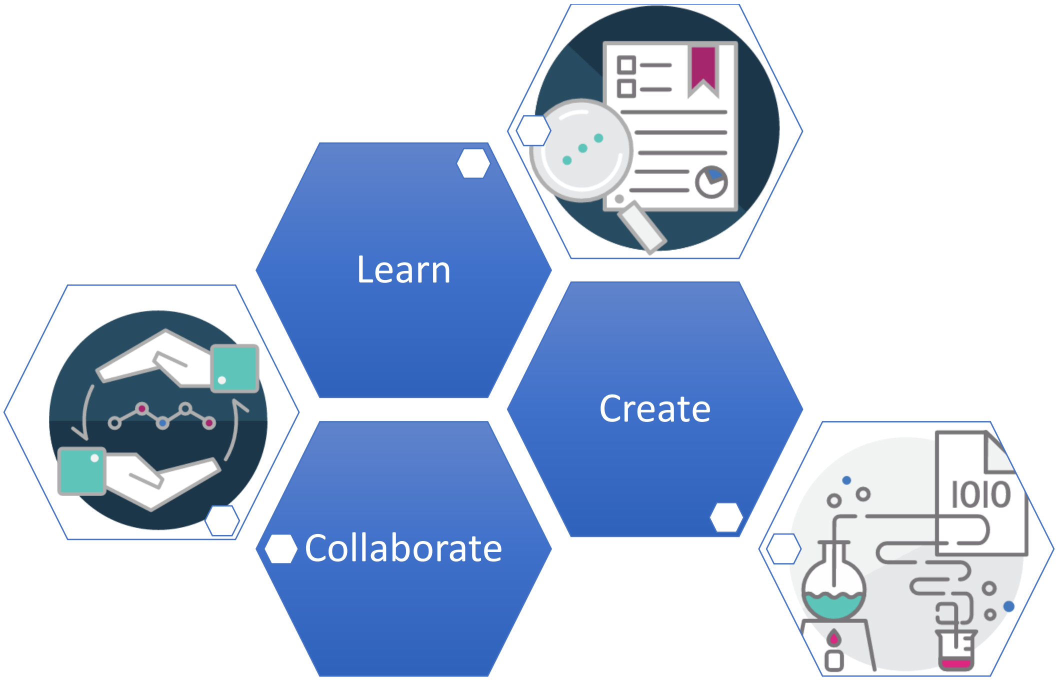
IBM Cloud

Data Science & Machine Learning 101

Technical Boot Camp

Lab Guide



Data Science

Machine Learning

Notices and Disclaimers

The information contained in these materials is provided for informational purposes only, and is provided AS IS without warranty of any kind, express or implied. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, these materials. Nothing contained in these materials is intended to, nor shall have the effect of, creating any warranties or representations from IBM or its suppliers or licensors, or altering the terms and conditions of the applicable license agreement governing the use of IBM software. References in these materials to IBM products, programs, or services do not imply that they will be available in all countries in which IBM operates. This information is based on current IBM product plans and strategy, which are subject to change by IBM without notice. Product release dates and/or capabilities referenced in these materials may change at any time at IBM's sole discretion based on market opportunities or other factors, and are not intended to be a commitment to future product or feature availability in any way.

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

IBM, the IBM logo and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at “Copyright and trademark information” at ibm.com/legal/copytrade.shtml

Other company, product and service names may be trademarks or service marks of others

Document Revision History

Rev #	File Name	Date
1.0	Experiencing Data Science Boot Camp Lab Guide.docx	09/14/2017
1.1	Data Science & Machine Learning 101.docx	09/28/2017

Prepared & Revised by:
Louis Frolio – louis.frolio@ibm.com
Ashley Troggio – atroggio@us.ibm.com
Dallas Sinnett - dsinnett@us.ibm.com
Darrel Pyle - darrel.pyle@ibm.com

Table of Contents

Lab Environment Overview

Installed Software and Tools

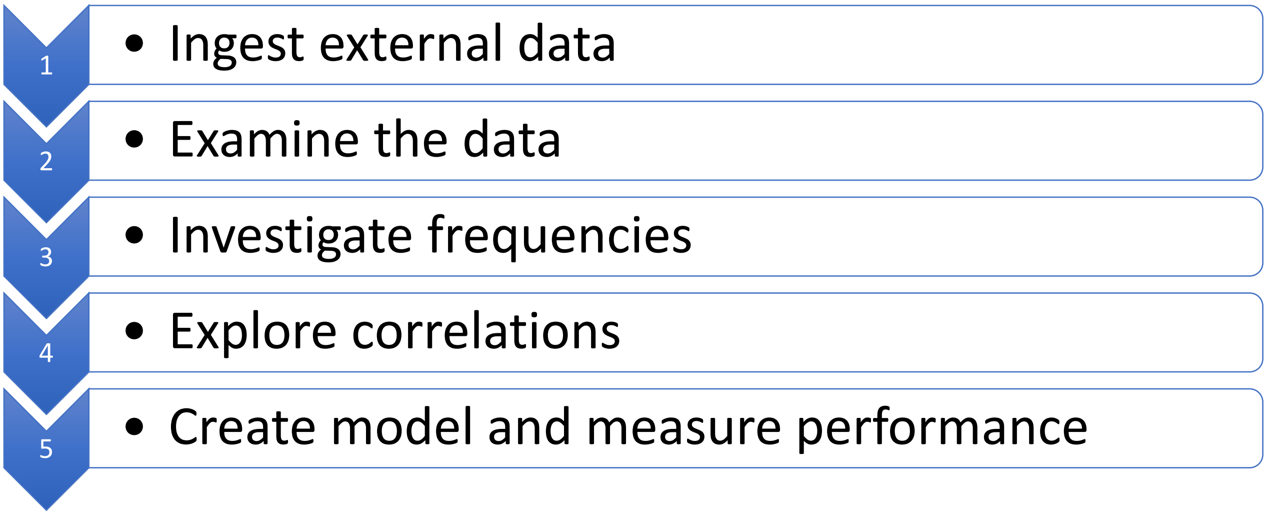
Software	Link
IBM Data Science Experience (DSX)	https://datascience.ibm.com/
IBM SPSS Statistics	http://www-03.ibm.com/software/products/no/spss-stats-base
Jupyter	http://jupyter.org/
GitHub	https:/github.org/
Anaconda	https://www.anaconda.com/
RStudio	https://www.rstudio.com/

Module 2: Statistics

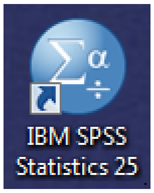
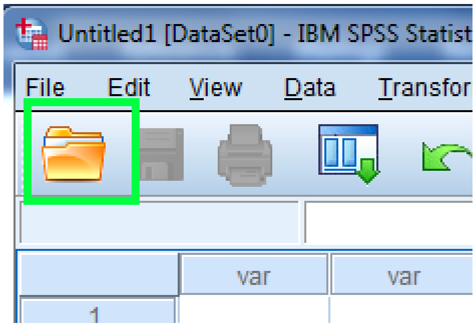
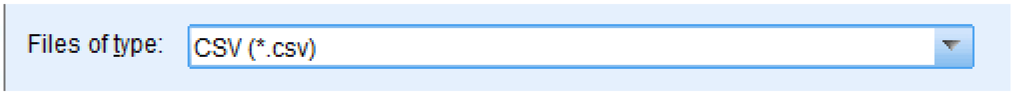
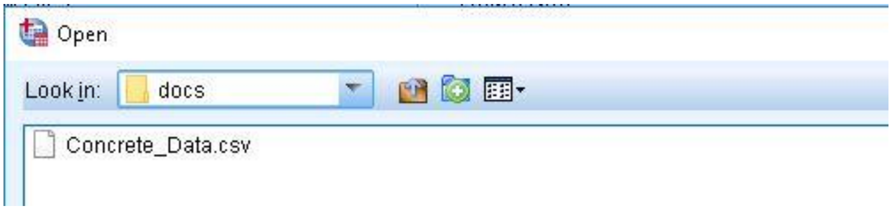
Purpose:	This lab introduces the subject of statistics and the process of performing statistical analysis. After completing the lab, you should be able to: • Ingest an external data into IBM SPSS Statistics • Explore the characteristics of the dataset • Examine its descriptive statistics • Create a statistical model

Tasks:	Tasks you will complete in this lab exercise include: • Load data • Exploratory Analysis ○ Analyze the data using visualizations ○ Test the data for correlations ○ Create a statistical model ○ Measure model performance

Module 2: Lab Workflow Overview



1. Ingest external data
2. Examine the data
3. Investigate frequencies
4. Explore correlations
5. Create model and measure performance

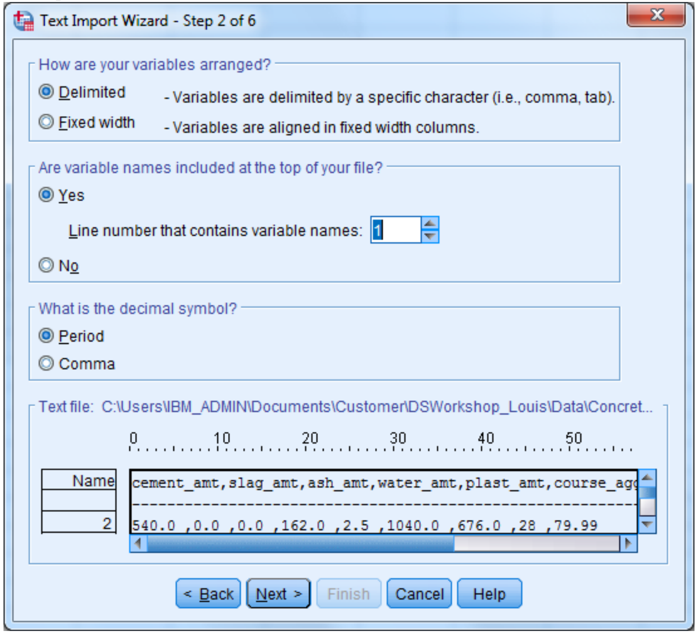
Step	Action
1	**Ingest external data** a. Double-click the **IBM SPSS Statistics 25** icon.  b. Close the dialog box that appears. c. Click on the **Open Data** icon in the toolbar.  Untitled1 [DataSet0] - IBM SPSS Statist File Edit View Data Transfor var var 1 d. Change the file type, in the middle of the dialog box, to **CSV** by clicking in the dropdown box.  Files of type: CSV (*.csv) e. At the top of the dialog box, navigate to the **Concrete_Data.CSV** file. C:\Data Science Bootcamp\docs  Look in: docs Concrete_Data.csv f. Click on file name and then **Open**.

Step	Action

g. There are many options when reading a flat file. You will be lead through 6 steps to ensure the data is properly read. Just click **Next** on Step 1.

In step 2, 3 and 4, match the selections below.

Step 2



Text Import Wizard - Step 2 of 6

How are your variables arranged?

● Delimited - Variables are delimited by a specific character (i.e., comma, tab).

○ Fixed width - Variables are aligned in fixed width columns.

Are variable names included at the top of your file?

● Yes

Line number that contains variable names: 1

○ No

What is the decimal symbol?

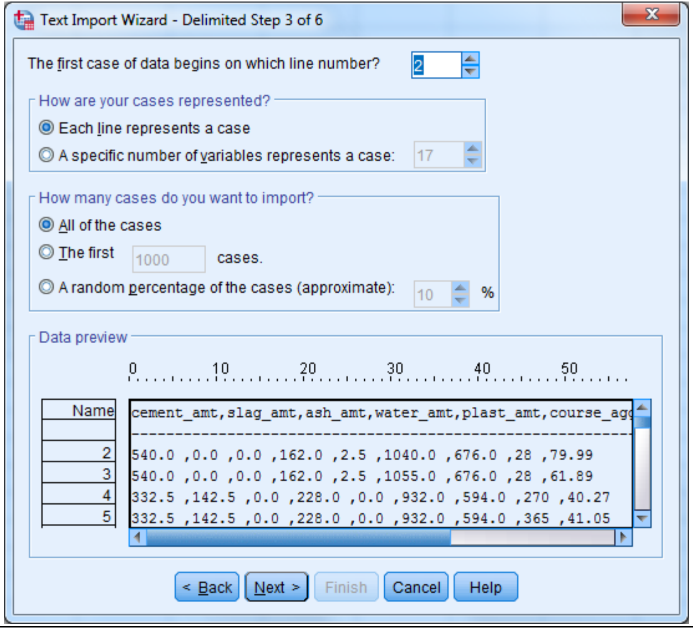
● Period

○ Comma

Text file: C:\Users\IBM_ADMIN\Documents\Customer\DSWorkshop_Louis\Data\Concret...

```
Name  cement_amt,slag_amt,ash_amt,water_amt,plast_amt,course_agg
      ---------------------------------------------------------
   2  540.0 ,0.0 ,0.0 ,162.0 ,2.5 ,1040.0 ,676.0 ,28 ,79.99
```

Step 3



Text Import Wizard - Delimited Step 3 of 6

The first case of data begins on which line number? 2

How are your cases represented?

● Each line represents a case

○ A specific number of variables represents a case: 17

How many cases do you want to import?

● All of the cases

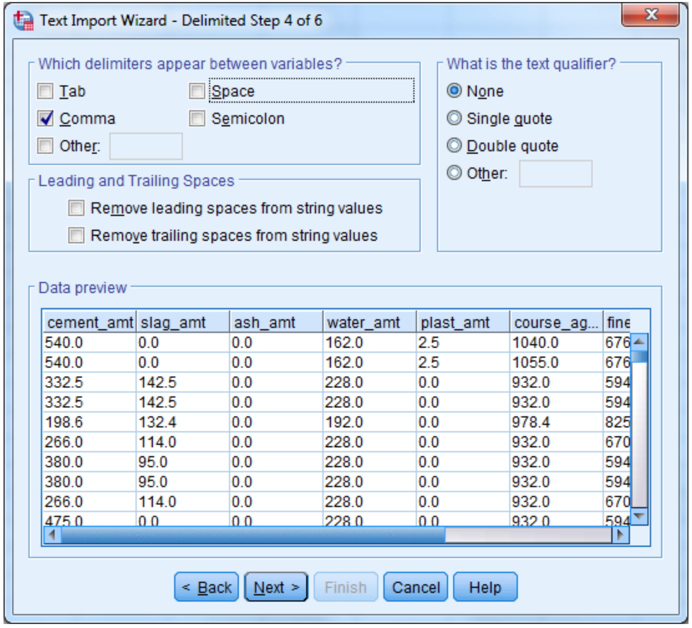
○ The first 1000 cases.

○ A random percentage of the cases (approximate): 10 %

Data preview

```
Name  cement_amt,slag_amt,ash_amt,water_amt,plast_amt,course_agg
      ---------------------------------------------------------
   2  540.0 ,0.0 ,0.0 ,162.0 ,2.5 ,1040.0 ,676.0 ,28 ,79.99
   3  540.0 ,0.0 ,0.0 ,162.0 ,2.5 ,1055.0 ,676.0 ,28 ,61.89
   4  332.5 ,142.5 ,0.0 ,228.0 ,0.0 ,932.0 ,594.0 ,270 ,40.27
   5  332.5 ,142.5 ,0.0 ,228.0 ,0.0 ,932.0 ,594.0 ,365 ,41.05
```

Step	Action
	Step 4



Text Import Wizard - Delimited Step 4 of 6

Which delimiters appear between variables?

☐ Tab ☐ Space
☑ Comma ☐ Semicolon
☐ Other:

What is the text qualifier?

● None
○ Single quote
○ Double quote
○ Other:

Leading and Trailing Spaces

☐ Remove leading spaces from string values
☐ Remove trailing spaces from string values

Data preview

cement_amt	slag_amt	ash_amt	water_amt	plast_amt	course_ag...	fine
540.0	0.0	0.0	162.0	2.5	1040.0	676
540.0	0.0	0.0	162.0	2.5	1055.0	676
332.5	142.5	0.0	228.0	0.0	932.0	594
332.5	142.5	0.0	228.0	0.0	932.0	594
198.6	132.4	0.0	192.0	0.0	978.4	825
266.0	114.0	0.0	228.0	0.0	932.0	670
380.0	95.0	0.0	228.0	0.0	932.0	594
380.0	95.0	0.0	228.0	0.0	932.0	594
266.0	114.0	0.0	228.0	0.0	932.0	670
475.0	0.0	0.0	228.0	0.0	932.0	594

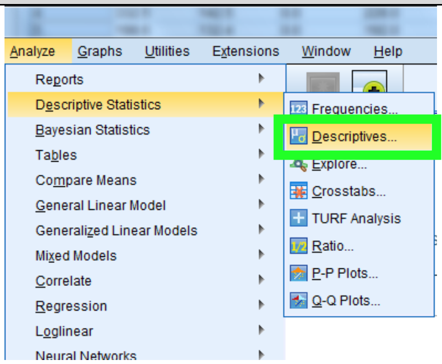
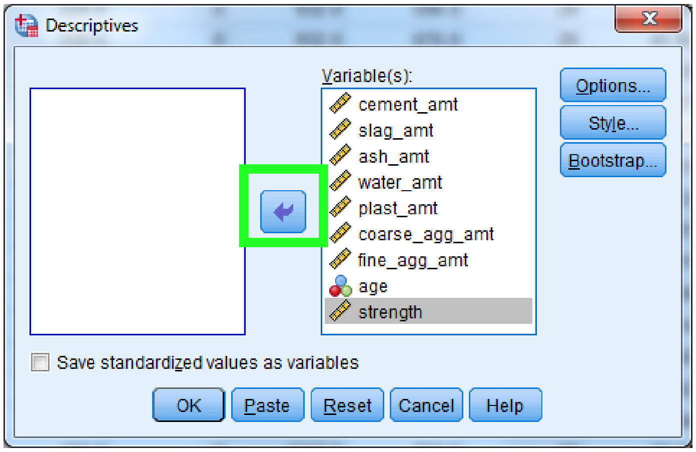
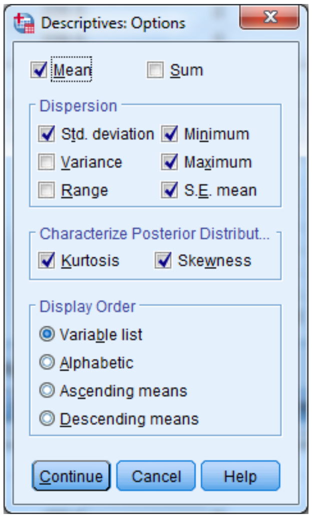
In step 5, click **Next**. In step 6, click **Finish**.

h. A new window, called the Output window, appears and then moves to the background. This records all activities as the product is used and is where the results of the procedures performed on the data will be shown.

i. It is helpful to know that SPSS can read many data formats. You can get a sense for this by exploring the various options for opening data files in the File menu.

2 **Examine the data**

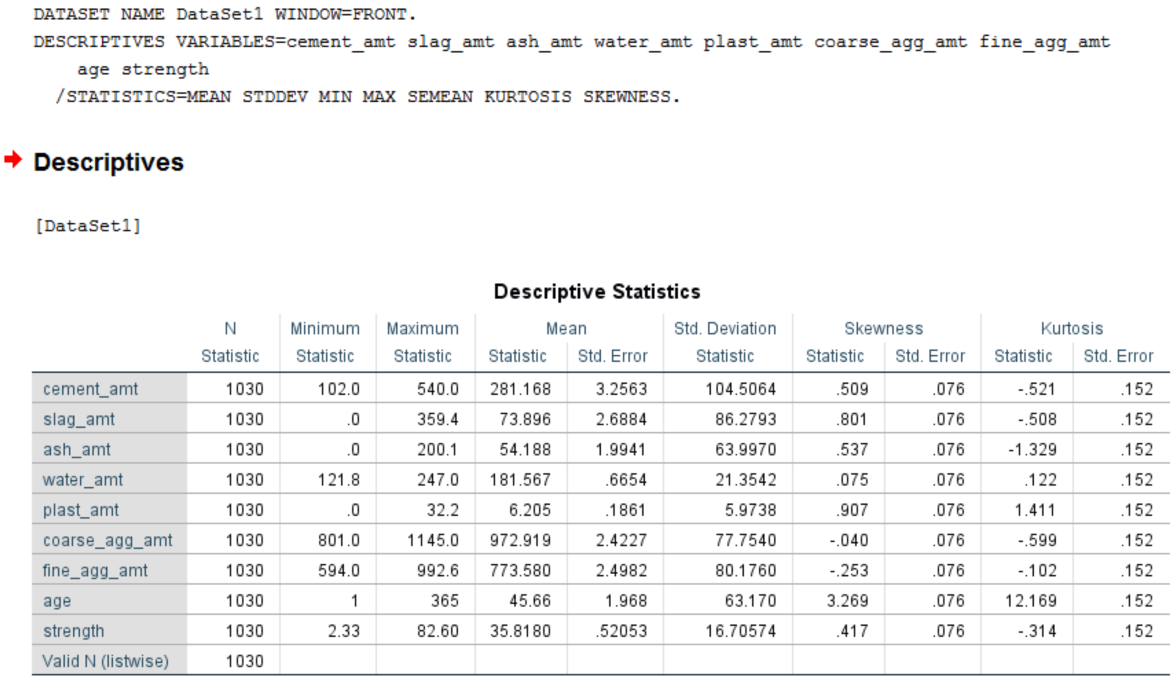
a. Click the Analyze menu in the menu bar. Hover over Descriptive Statistics and click **Descriptives** in the sub-menu.

Step	Action
	 Analyze Graphs Utilities Extensions Window Help Reports Descriptive Statistics Bayesian Statistics Tables Compare Means General Linear Model Generalized Linear Models Mixed Models Correlate Regression Loglinear Neural Networks Frequencies... Descriptives... Explore... Crosstabs... TURF Analysis Ratio... P-P Plots... Q-Q Plots... b. Select all the variables; then, click the arrow box to move all the variables to the Variable(s) box on the right.  Descriptives Options... Style... Bootstrap... OK Paste Reset Cancel Help Variable(s): cement_amt slag_amt ash_amt water_amt plast_amt coarse_agg_amt fine_agg_amt age strength ☐ Save standardized values as variables c. Click **Options**. Make the selections as indicated below.  ☑ Mean ☐ Sum Dispersion ☑ Std. deviation ☑ Minimum ☐ Variance ☑ Maximum ☐ Range ☑ S.E. mean Characterize Posterior Distribut... ☑ Kurtosis ☑ Skewness Display Order ● Variable list ○ Alphabetic ○ Ascending means ○ Descending means

Step	Action

d. Click **Continue**. Then, click **OK** in the Descriptives dialog.

e. Examine the statistics about each variable.



```
DATASET NAME DataSet1 WINDOW=FRONT.
DESCRIPTIVES VARIABLES=cement_amt slag_amt ash_amt water_amt plast_amt coarse_agg_amt fine_agg_amt
    age strength
  /STATISTICS=MEAN STDDEV MIN MAX SEMEAN KURTOSIS SKEWNESS.
```

➜ **Descriptives**

```
[DataSet1]
```

Descriptive Statistics

	N	Minimum	Maximum	Mean		Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Std. Error	Statistic	Std. Error
cement_amt	1030	102.0	540.0	281.168	3.2563	104.5064	.509	.076	-.521	.152
slag_amt	1030	.0	359.4	73.896	2.6884	86.2793	.801	.076	-.508	.152
ash_amt	1030	.0	200.1	54.188	1.9941	63.9970	.537	.076	-1.329	.152
water_amt	1030	121.8	247.0	181.567	.6654	21.3542	.075	.076	.122	.152
plast_amt	1030	.0	32.2	6.205	.1861	5.9738	.907	.076	1.411	.152
coarse_agg_amt	1030	801.0	1145.0	972.919	2.4227	77.7540	-.040	.076	-.599	.152
fine_agg_amt	1030	594.0	992.6	773.580	2.4982	80.1760	-.253	.076	-.102	.152
age	1030	1	365	45.66	1.968	63.170	3.269	.076	12.169	.152
strength	1030	2.33	82.60	35.8180	.52053	16.70574	.417	.076	-.314	.152
Valid N (listwise)	1030									

As shown above this table, there is SPSS generated code.

Do you see terms that you recognize?________________________

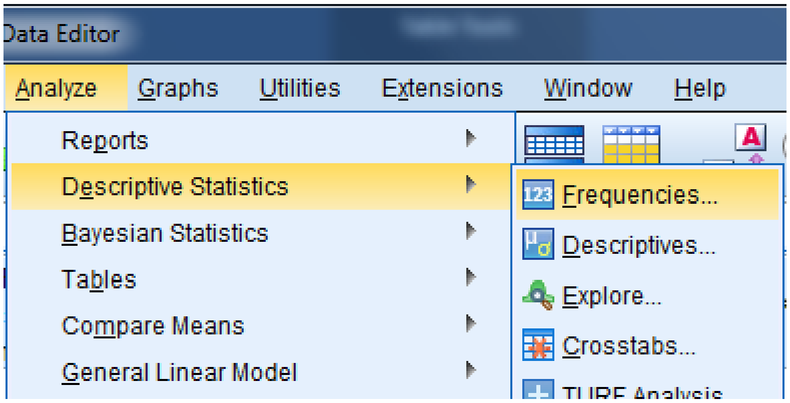
What does the code mean? ________________________________

Is there a relationship between the mean, standard deviation and skewness? If yes, explain.

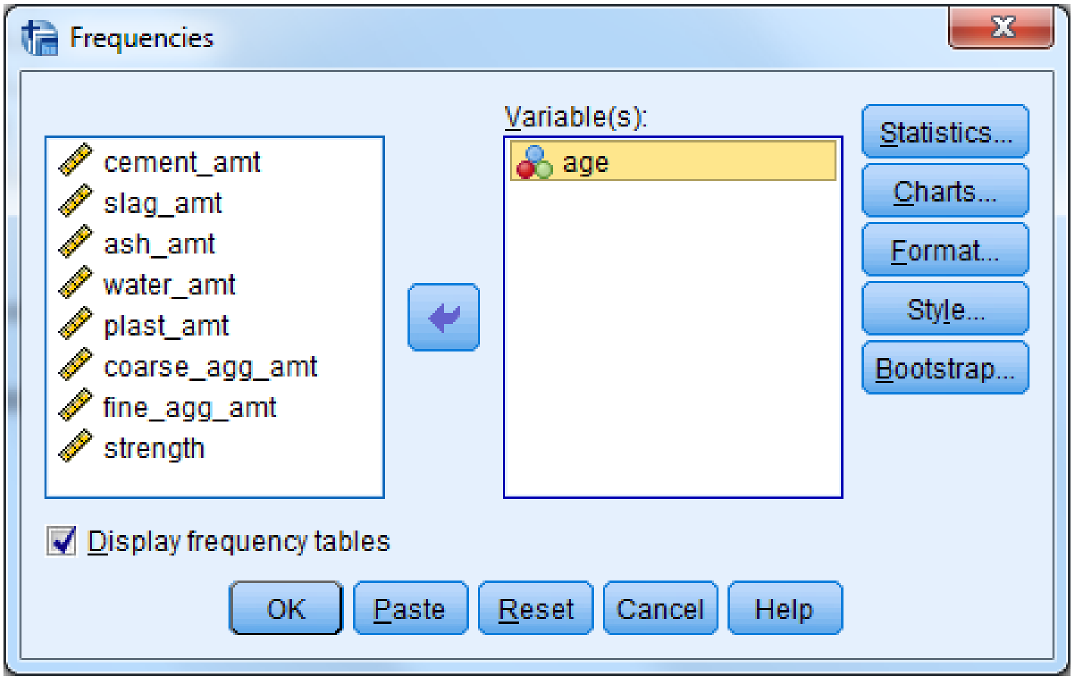
____

Step	Action
3	**Investigate frequencies**

a. Click the Analyze menu in the menu bar. Hover over Descriptive Statistics and click **Frequencies** in the sub-menu.

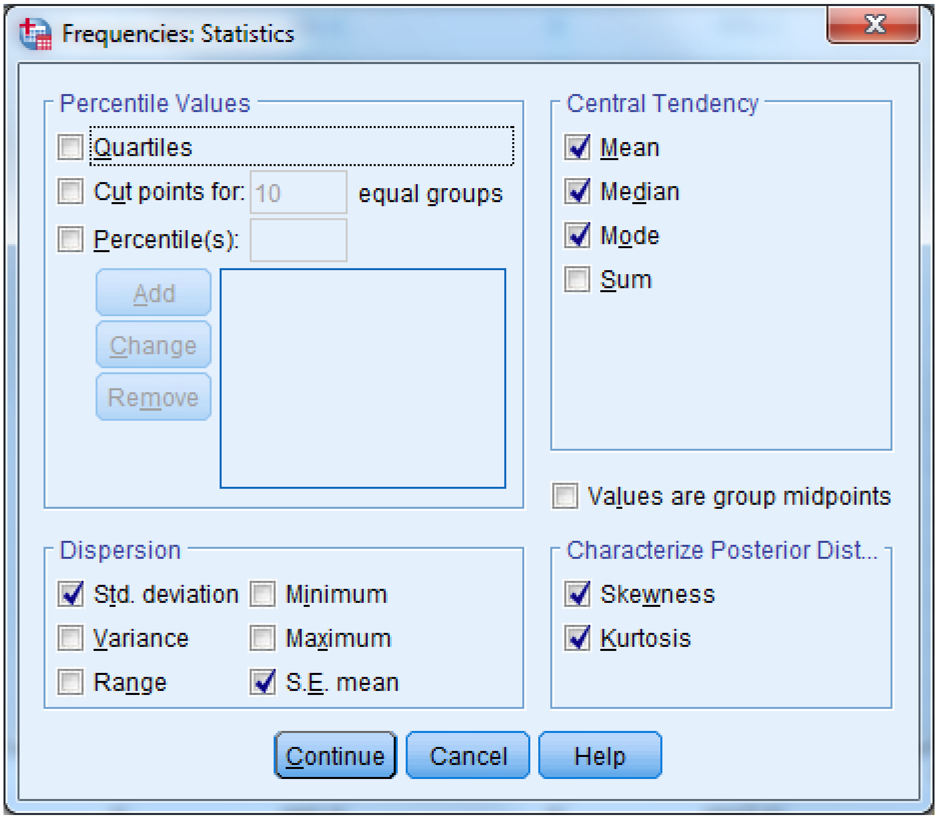
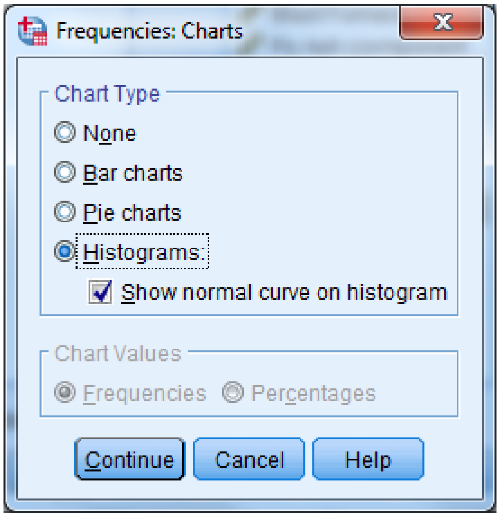


b. Move the Age variable to the Variable(s) box. Ensure the ‘Display frequency tables’ box is checked.



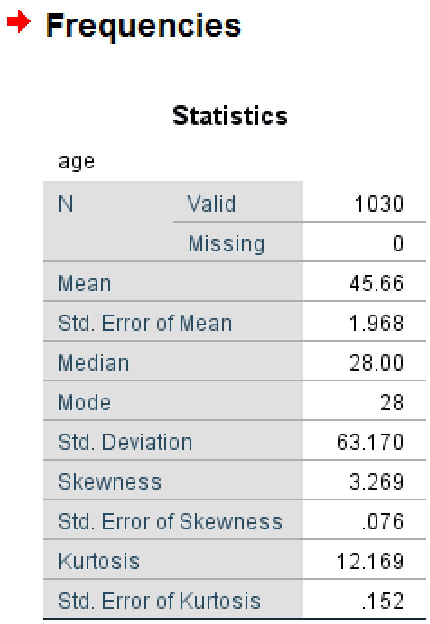
Variable(s): age

☑ Display frequency tables

Step	Action
	c. Click **Statistics** and match the selections below.  Frequencies: Statistics Percentile Values ☐ Quartiles ☐ Cut points for: 10 equal groups ☐ Percentile(s): Central Tendency ☑ Mean ☑ Median ☑ Mode ☐ Sum ☐ Values are group midpoints Dispersion ☑ Std. deviation ☐ Minimum ☐ Variance ☐ Maximum ☐ Range ☑ S.E. mean Characterize Posterior Dist... ☑ Skewness ☑ Kurtosis d. Click **Continue**, then **Charts**. Match the selections below.  Frequencies: Charts Chart Type ○ None ○ Bar charts ○ Pie charts ● Histograms: ☑ Show normal curve on histogram Chart Values ● Frequencies ○ Percentages

Step	Action

e. Click **Continue**; then, **OK**. You will see the descriptive statistics; the same information you saw in the previous exercise.



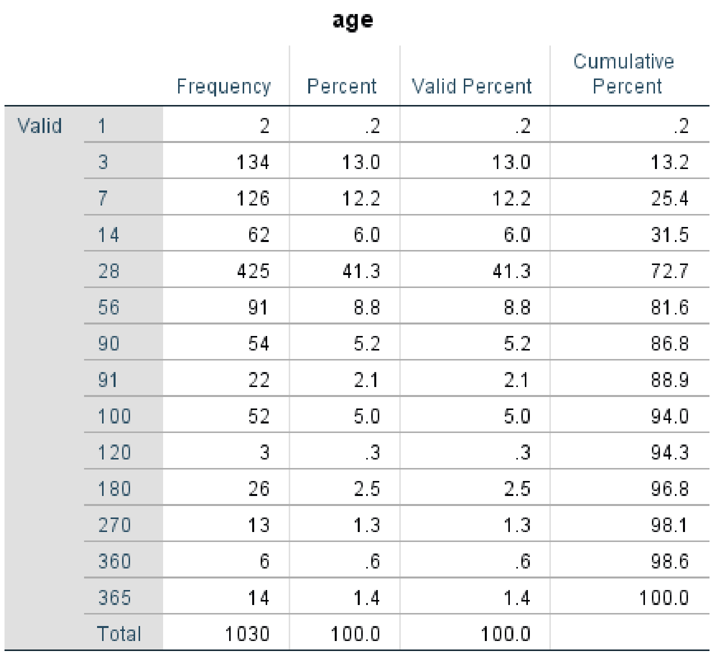
Frequencies

Statistics

age

N	Valid	1030
	Missing	0
Mean		45.66
Std. Error of Mean		1.968
Median		28.00
Mode		28
Std. Deviation		63.170
Skewness		3.269
Std. Error of Skewness		.076
Kurtosis		12.169
Std. Error of Kurtosis		.152

f. Followed by the frequency table.

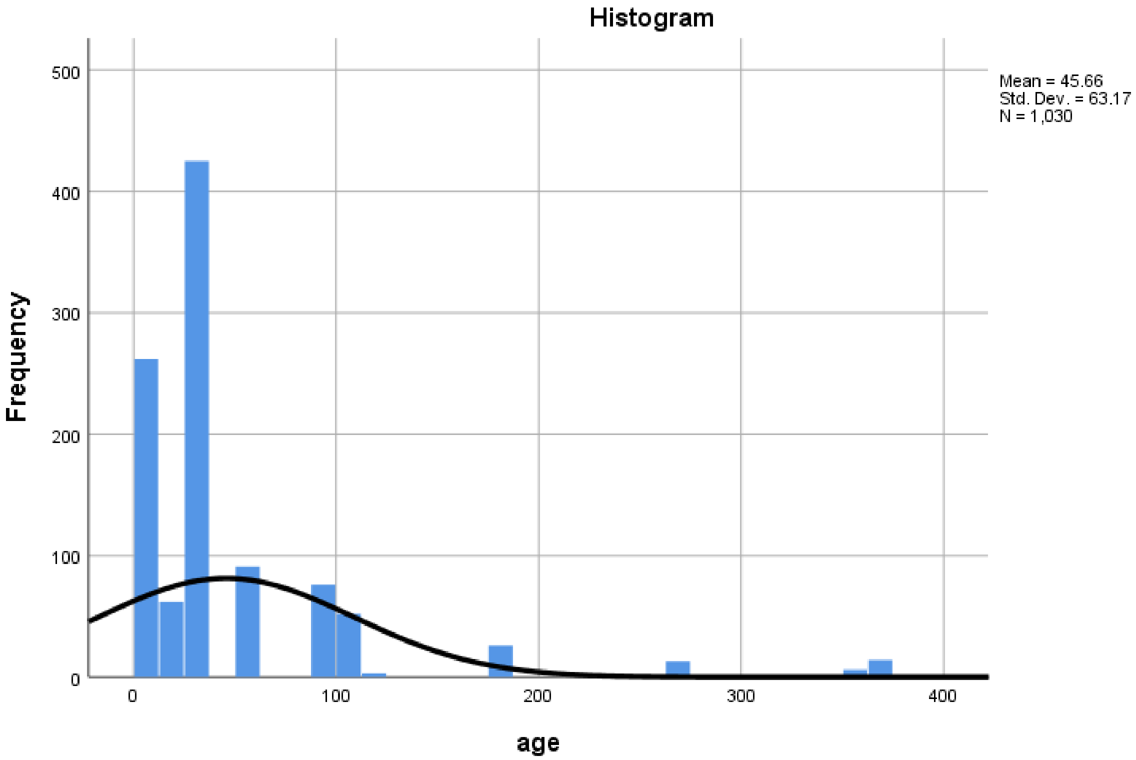


age

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	2	.2	.2	.2
	3	134	13.0	13.0	13.2
	7	126	12.2	12.2	25.4
	14	62	6.0	6.0	31.5
	28	425	41.3	41.3	72.7
	56	91	8.8	8.8	81.6
	90	54	5.2	5.2	86.8
	91	22	2.1	2.1	88.9
	100	52	5.0	5.0	94.0
	120	3	.3	.3	94.3
	180	26	2.5	2.5	96.8
	270	13	1.3	1.3	98.1
	360	6	.6	.6	98.6
	365	14	1.4	1.4	100.0
	Total	1030	100.0	100.0	

Step	Action

g. This is followed by a histogram of the variable.

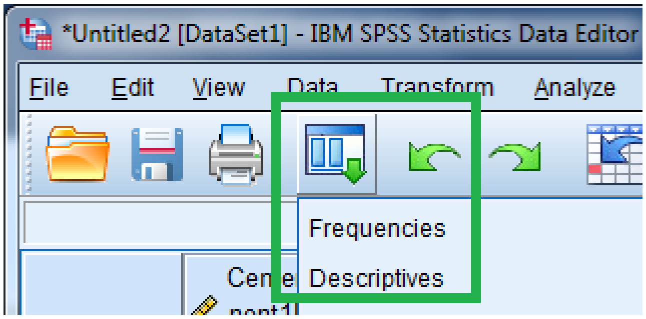
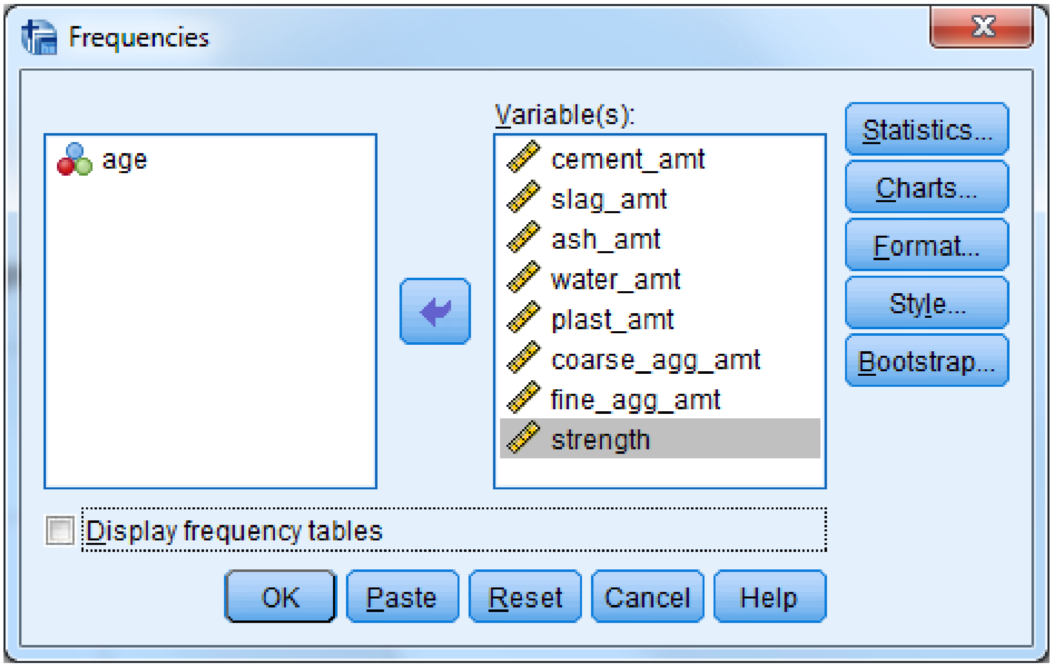


h. The histogram includes a curved line that represents a normal distribution of the variable. It appears that this variable is not normally distributed. See the descriptive information and discuss the information as it relates to the normality of the data.

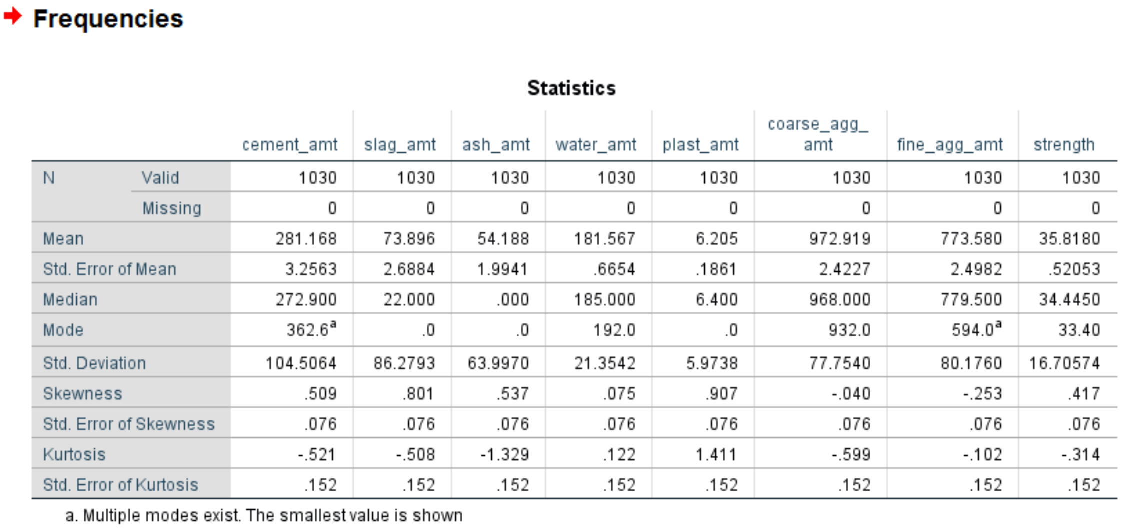
Think on this:
What do you notice about the Age variable?
How does this relate to the descriptive statistics we discussed earlier?

____

____

Step	Action
	i. Use the dialog recall button in the toolbar to bring up the Frequencies dialog again.  *Untitled2 [DataSet1] - IBM SPSS Statistics Data Editor File Edit View Data Transform Analyze Frequencies Descriptives j. Move the Age variable out of the Variable(s) box; and add all the other variables. Make sure to uncheck the box in the lower left. **Think on this:** Do you know why this is a good thing to do? Leave it checked if you need some insight.  Frequencies Statistics... Charts... Format... Style... Bootstrap... OK Paste Reset Cancel Help age Variable(s): cement_amt slag_amt ash_amt water_amt plast_amt coarse_agg_amt fine_agg_amt strength □ Display frequency tables

Step	Action
	k. Click **OK.** You see the frequency statistics for each variable.



Frequencies

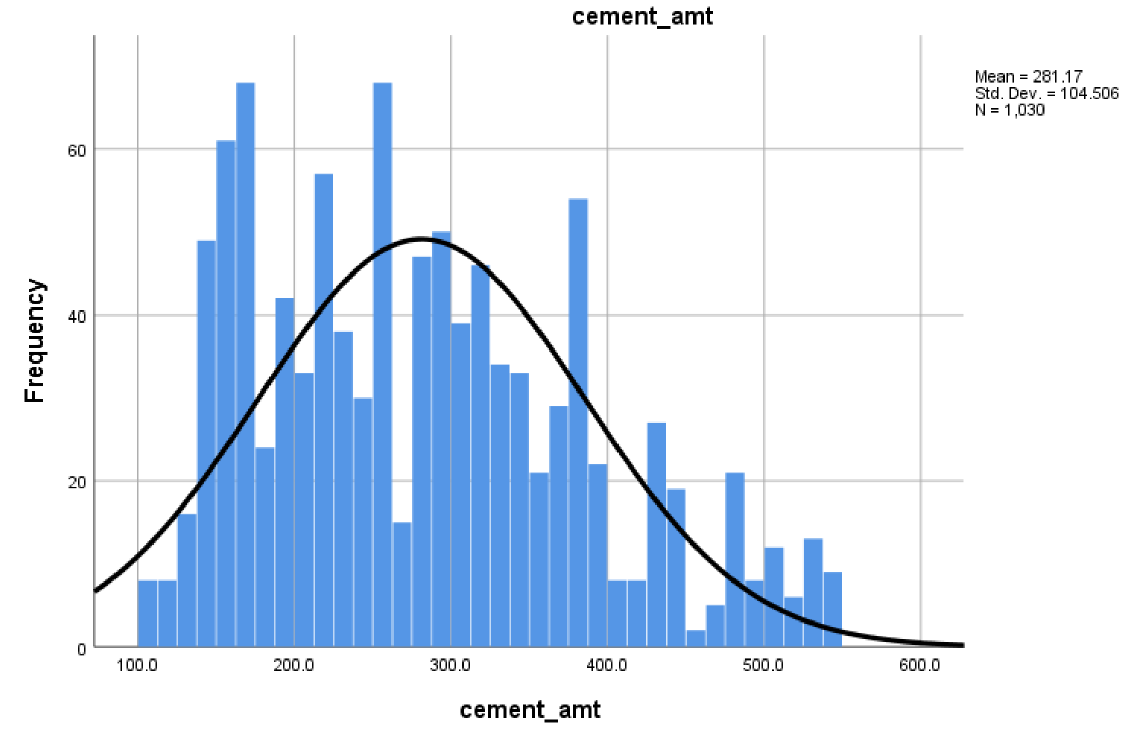
Statistics

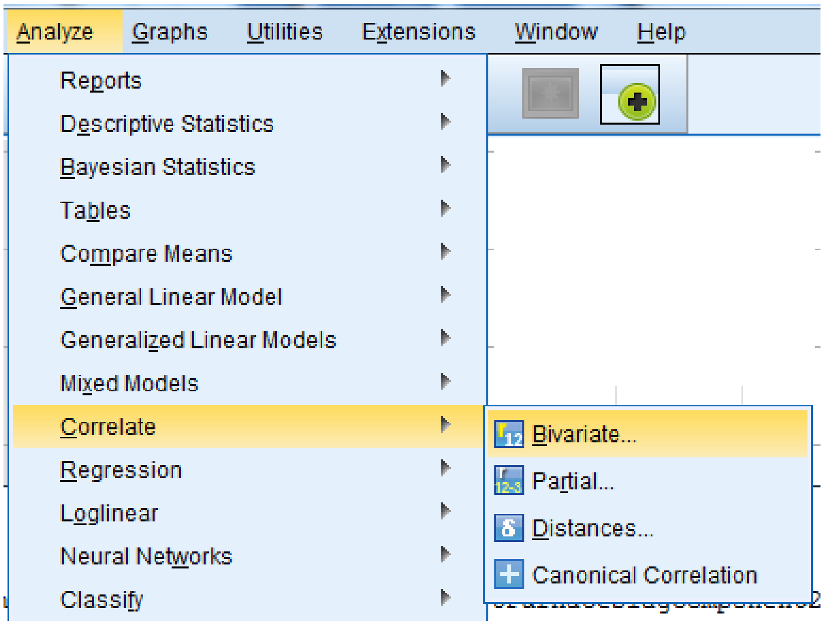
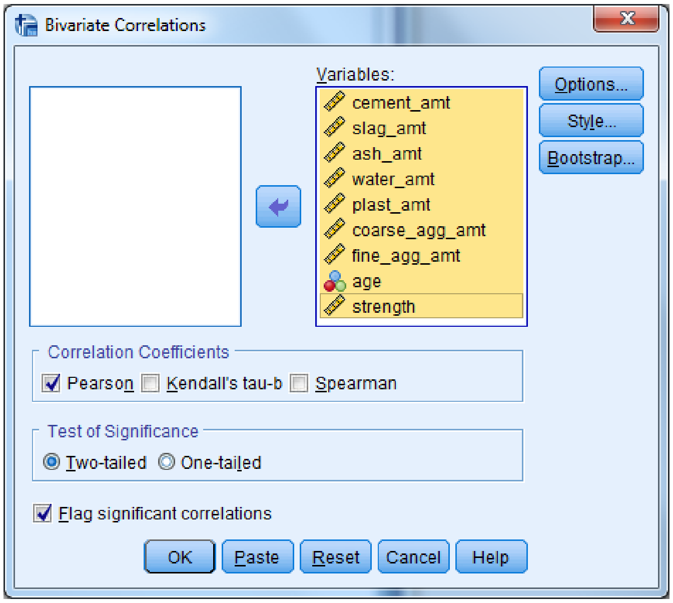
		cement_amt	slag_amt	ash_amt	water_amt	plast_amt	coarse_agg_amt	fine_agg_amt	strength
N	Valid	1030	1030	1030	1030	1030	1030	1030	1030
	Missing	0	0	0	0	0	0	0	0
Mean		281.168	73.896	54.188	181.567	6.205	972.919	773.580	35.8180
Std. Error of Mean		3.2563	2.6884	1.9941	.6654	.1861	2.4227	2.4982	.52053
Median		272.900	22.000	.000	185.000	6.400	968.000	779.500	34.4450
Mode		362.6[a]	.0	.0	192.0	.0	932.0	594.0[a]	33.40
Std. Deviation		104.5064	86.2793	63.9970	21.3542	5.9738	77.7540	80.1760	16.70574
Skewness		.509	.801	.537	.075	.907	-.040	-.253	.417
Std. Error of Skewness		.076	.076	.076	.076	.076	.076	.076	.076
Kurtosis		-.521	-.508	-1.329	.122	1.411	-.599	-.102	-.314
Std. Error of Kurtosis		.152	.152	.152	.152	.152	.152	.152	.152

a. Multiple modes exist. The smallest value is shown

Followed by a histogram for each variable as a visual overview. Here is an example.

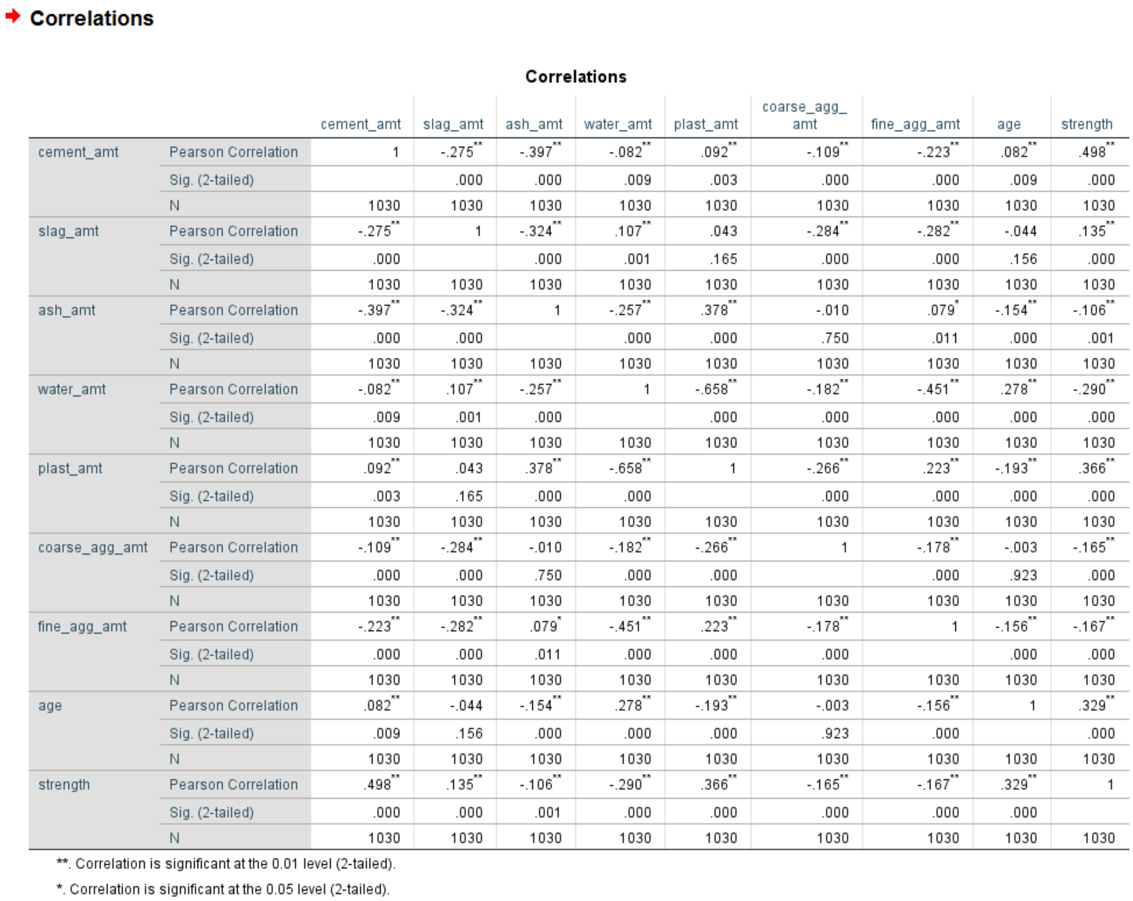
Histogram



Step	Action
4	**Explore correlations** Next, explore the relationships between different variables. a. In the Analyze menu, go down to the Correlate entry and click on **Bivariate**.  Analyze Graphs Utilities Extensions Window Help Reports Descriptive Statistics Bayesian Statistics Tables Compare Means General Linear Model Generalized Linear Models Mixed Models Correlate Regression Loglinear Neural Networks Classify Bivariate... Partial... Distances... Canonical Correlation b. Move all the variables to the Variables box and select the options as shown:  Bivariate Correlations Variables: cement_amt slag_amt ash_amt water_amt plast_amt coarse_agg_amt fine_agg_amt age strength Correlation Coefficients ☑ Pearson □ Kendall's tau-b □ Spearman Test of Significance ● Two-tailed ○ One-tailed ☑ Flag significant correlations Options... Style... Bootstrap... OK Paste Reset Cancel Help

Step	Action

Click **OK.** You will see the following correlation table.



Correlations

Correlations

		cement_amt	slag_amt	ash_amt	water_amt	plast_amt	coarse_agg_ amt	fine_agg_amt	age	strength
cement_amt	Pearson Correlation	1	-.275**	-.397**	-.082**	.092**	-.109**	-.223**	.082**	.498**
	Sig. (2-tailed)		.000	.000	.009	.003	.000	.000	.009	.000
	N	1030	1030	1030	1030	1030	1030	1030	1030	1030
slag_amt	Pearson Correlation	-.275**	1	-.324**	.107**	.043	-.284**	-.282**	-.044	.135**
	Sig. (2-tailed)	.000		.000	.001	.165	.000	.000	.156	.000
	N	1030	1030	1030	1030	1030	1030	1030	1030	1030
ash_amt	Pearson Correlation	-.397**	-.324**	1	-.257**	.378**	-.010	.079*	-.154**	-.106**
	Sig. (2-tailed)	.000	.000		.000	.000	.750	.011	.000	.001
	N	1030	1030	1030	1030	1030	1030	1030	1030	1030
water_amt	Pearson Correlation	-.082**	.107**	-.257**	1	-.658**	-.182**	-.451**	.278**	-.290**
	Sig. (2-tailed)	.009	.001	.000		.000	.000	.000	.000	.000
	N	1030	1030	1030	1030	1030	1030	1030	1030	1030
plast_amt	Pearson Correlation	.092**	.043	.378**	-.658**	1	-.266**	.223**	-.193**	.366**
	Sig. (2-tailed)	.003	.165	.000	.000		.000	.000	.000	.000
	N	1030	1030	1030	1030	1030	1030	1030	1030	1030
coarse_agg_amt	Pearson Correlation	-.109**	-.284**	-.010	-.182**	-.266**	1	-.178**	-.003	-.165**
	Sig. (2-tailed)	.000	.000	.750	.000	.000		.000	.923	.000
	N	1030	1030	1030	1030	1030	1030	1030	1030	1030
fine_agg_amt	Pearson Correlation	-.223**	-.282**	.079*	-.451**	.223**	-.178**	1	-.156**	-.167**
	Sig. (2-tailed)	.000	.000	.011	.000	.000	.000		.000	.000
	N	1030	1030	1030	1030	1030	1030	1030	1030	1030
age	Pearson Correlation	.082**	-.044	-.154**	.278**	-.193**	-.003	-.156**	1	.329**
	Sig. (2-tailed)	.009	.156	.000	.000	.000	.923	.000		.000
	N	1030	1030	1030	1030	1030	1030	1030	1030	1030
strength	Pearson Correlation	.498**	.135**	-.106**	-.290**	.366**	-.165**	-.167**	.329**	1
	Sig. (2-tailed)	.000	.000	.001	.000	.000	.000	.000	.000	
	N	1030	1030	1030	1030	1030	1030	1030	1030	1030

**. Correlation is significant at the 0.01 level (2-tailed).

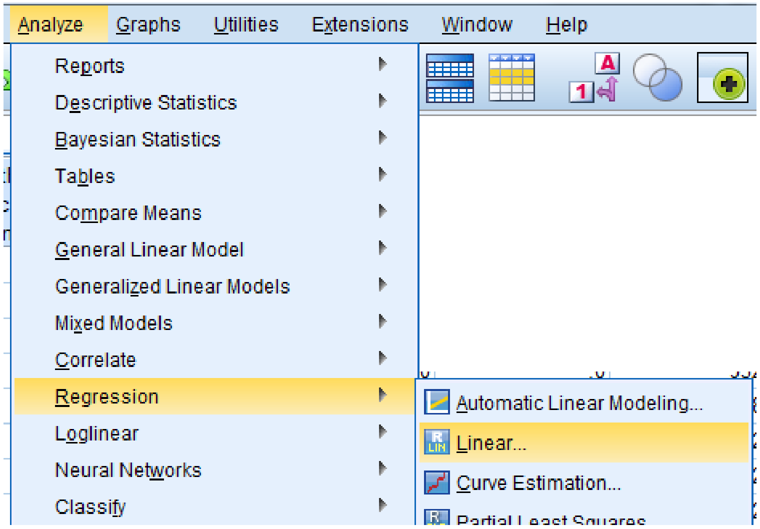
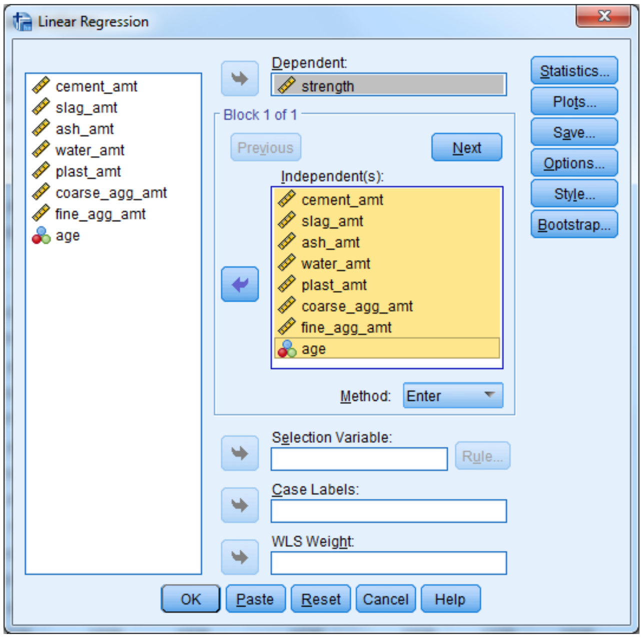
*. Correlation is significant at the 0.05 level (2-tailed).

Scan the rows labelled ‘Pearson Correlation’.
What is the largest positive correlation coefficient? _______________
What is the largest negative correlation coefficient? _________________
What are some of the significant correlations? ________________
Do any of them surprise you? __________________

5

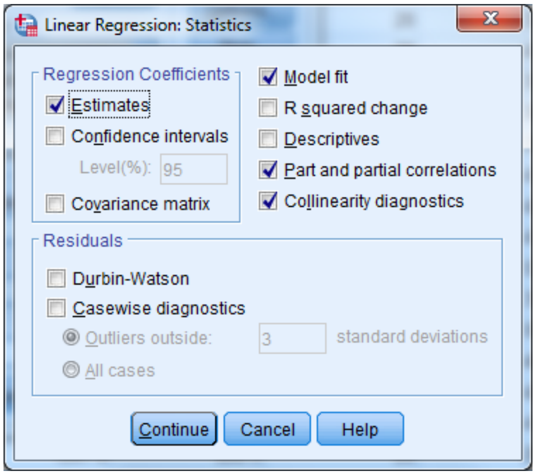
Create model and measure performance

Creating a model of the data (a mathematical representation of what goes on in the data) is a goal in data science. A model can be used to gain insights from the data, but is more often viewed as a means to determine future outcomes. For example, you can get an idea of what the compressive strength of concrete would be if you varied the amount of one or more ingredients.

Step	Action
	In this exercise, a multiple linear regression model will be built and evaluated. a. In the Analyze menu, hover over the Regression select then click on **Linear** in the sub-menu.  Analyze Graphs Utilities Extensions Window Help Reports Descriptive Statistics Bayesian Statistics Tables Compare Means General Linear Model Generalized Linear Models Mixed Models Correlate Regression Loglinear Neural Networks Classify Automatic Linear Modeling... Linear... Curve Estimation... Partial Least Squares b. In the Linear Regression dialog, move 'Cement compressive strength' to the Dependent box and all the other variables to the Independent(s) box as shown.  Linear Regression cement_amt slag_amt ash_amt water_amt plast_amt coarse_agg_amt fine_agg_amt age Dependent: strength Block 1 of 1 Independent(s): cement_amt, slag_amt, ash_amt, water_amt, plast_amt, coarse_agg_amt, fine_agg_amt, age Method: Enter Selection Variable: Case Labels: WLS Weight: Statistics... Plots... Save... Options... Style... Bootstrap... Previous Next Rule... OK Paste Reset Cancel Help

Step	Action

c. Click **Statistics** and make the selections as shown below.



Linear Regression: Statistics

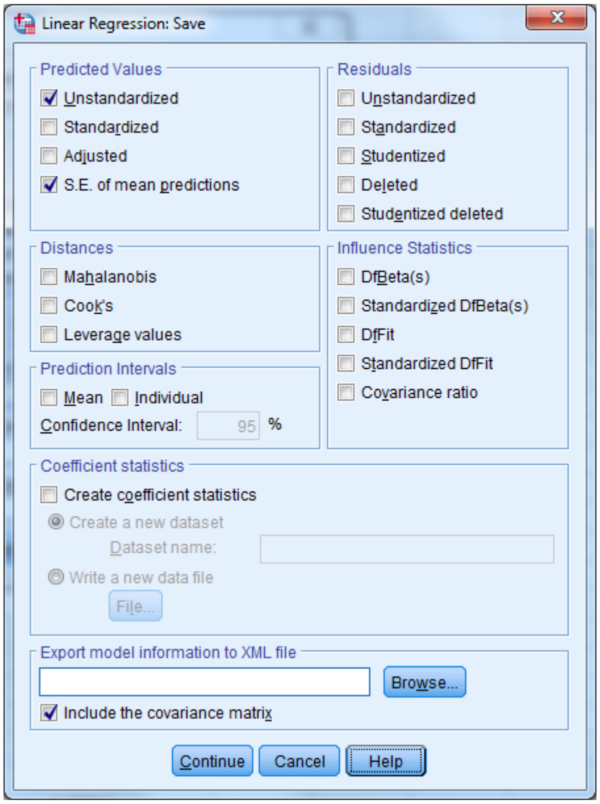
Regression Coefficients
☑ Estimates
□ Confidence intervals
Level(%): 95
□ Covariance matrix

☑ Model fit
□ R squared change
□ Descriptives
☑ Part and partial correlations
☑ Collinearity diagnostics

Residuals
□ Durbin-Watson
□ Casewise diagnostics
● Outliers outside: 3 standard deviations
○ All cases

d. Click **Continue.**

e. Click the **Save** button and make the selections as shown.



Linear Regression: Save

Predicted Values
☑ Unstandardized
□ Standardized
□ Adjusted
☑ S.E. of mean predictions

Residuals
□ Unstandardized
□ Standardized
□ Studentized
□ Deleted
□ Studentized deleted

Distances
□ Mahalanobis
□ Cook's
□ Leverage values

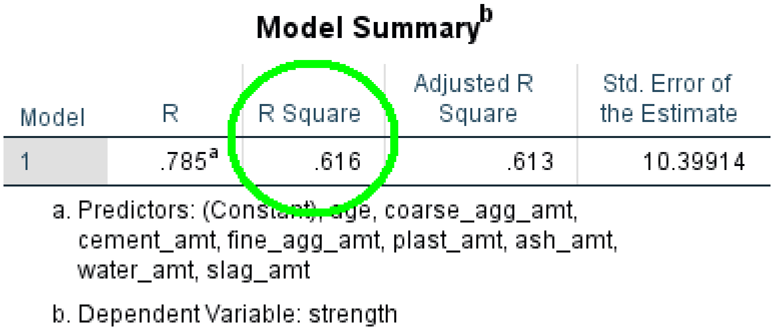
Influence Statistics
□ DfBeta(s)
□ Standardized DfBeta(s)
□ DfFit
□ Standardized DfFit
□ Covariance ratio

Prediction Intervals
□ Mean □ Individual
Confidence Interval: 95 %

Coefficient statistics
□ Create coefficient statistics
● Create a new dataset
Dataset name:
○ Write a new data file

Export model information to XML file
☑ Include the covariance matrix

Step	Action
	These selections add variables to the dataset to determine how well the model is performing. Click **Continue.**
	f. Click **OK;** then, review the results.
	g. Scroll down to the Model Summary. The R Square number indicates that about 61% of the variability in Compressive strength is explained by the model.



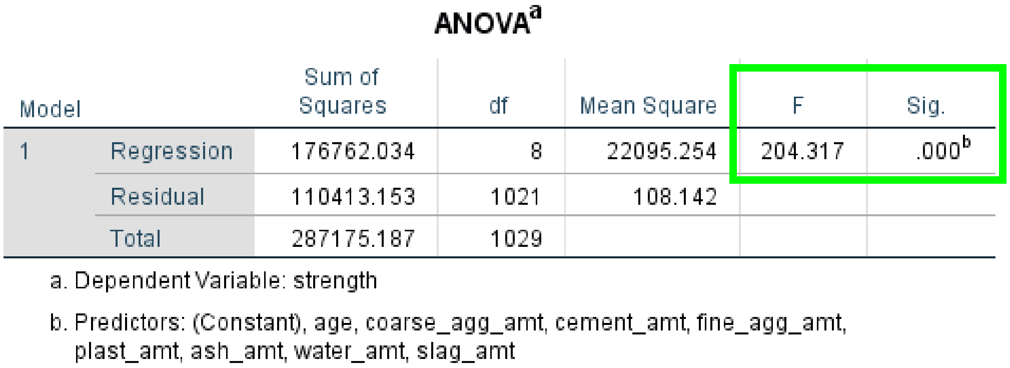
Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.785[a]	.616	.613	10.39914

a. Predictors: (Constant), age, coarse_agg_amt, cement_amt, fine_agg_amt, plast_amt, ash_amt, water_amt, slag_amt

b. Dependent Variable: strength

h. The next table in the Output window shows a significant F statistic, which means the model is better than guessing.



ANOVA[a]

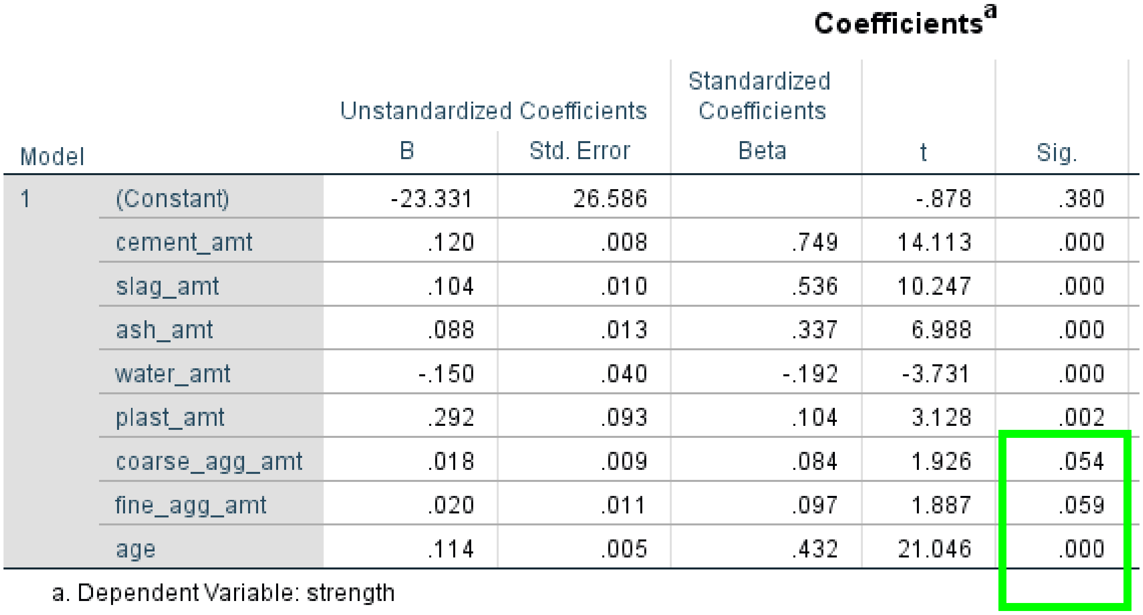
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	176762.034	8	22095.254	204.317	.000[b]
	Residual	110413.153	1021	108.142		
	Total	287175.187	1029			

a. Dependent Variable: strength

b. Predictors: (Constant), age, coarse_agg_amt, cement_amt, fine_agg_amt, plast_amt, ash_amt, water_amt, slag_amt

i. The model as a whole appears to be accurate at predicting the concrete strength. However, there are a few variables that don’t have a strong influence on the output.

Step	Action



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-23.331	26.586		-.878	.380
	cement_amt	.120	.008	.749	14.113	.000
	slag_amt	.104	.010	.536	10.247	.000
	ash_amt	.088	.013	.337	6.988	.000
	water_amt	-.150	.040	-.192	-3.731	.000
	plast_amt	.292	.093	.104	3.128	.002
	coarse_agg_amt	.018	.009	.084	1.926	.054
	fine_agg_amt	.020	.011	.097	1.887	.059
	age	.114	.005	.432	21.046	.000

a. Dependent Variable: strength

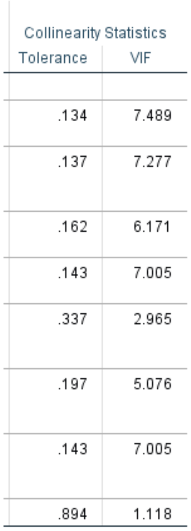
‘Coarse Aggregate’ and ‘Fine Aggregate’ are NOT significant. This indicates they contribute minimally to the model. It might be helpful to remove them from the model.

Think on this:
‘Age’ is also highlighted here even though it is significant. Consider the descriptive statistics from above.

Could Age have a negative impact on your analysis? Why or why not?

j. Scrolling to the right in the same table you see the Collinearity Statistics.

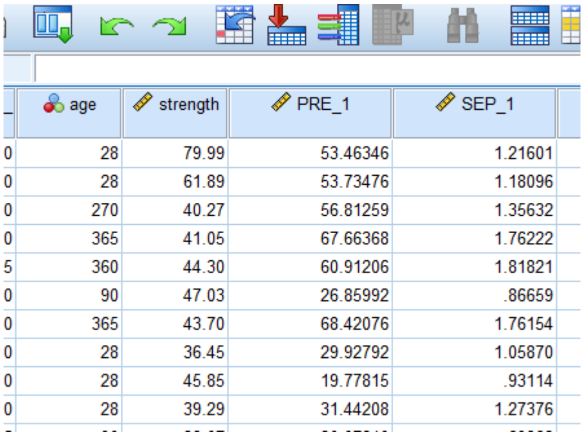
Step	Action



Collinearity Statistics	
Tolerance	VIF
.134	7.489
.137	7.277
.162	6.171
.143	7.005
.337	2.965
.197	5.076
.143	7.005
.894	1.118

Low tolerances, such as you see here, indicate high multicollinearity of the variables. Its associated Variable Inflation Factor (VIF) is considered problematic when it gets higher than 2. You see in the list that, except for Age, all are much higher than 2. So, our model might not be as useful as it could be.

k. Next, look at what the model actually did with the data. By saving the predicted values, they will be seen as a new variable in the dataset.



_	age	strength	PRE_1	SEP_1
0	28	79.99	53.46346	1.21601
0	28	61.89	53.73476	1.18096
0	270	40.27	56.81259	1.35632
0	365	41.05	67.66368	1.76222
5	360	44.30	60.91206	1.81821
0	90	47.03	26.85992	.86659
0	365	43.70	68.42076	1.76154
0	28	36.45	29.92792	1.05870
0	28	45.85	19.77815	.93114
0	28	39.29	31.44208	1.27376

PRE_1 is the variable where the model wrote the predicted value. That is, the model was built, then the data was scored by the model. SEP_1 is the Standard Error of the prediction. Comparing PRE_1 to Concrete compressive strength (the dependent variable) you begin to see that there are some big differences.

Step	Action
	Think on this: What further actions could be taken to refine the model? ____ ____

Module 2: Lab Summary

In this module, SPSS Statistics was used to examine the characteristics of individual variables within a dataset. An analysis was performed to better understand the shape and size of the data and discover relationships among the variables. Further evaluation was completed by using regression models in order to describe the relationships of the variables. Finally, the regression model’s performance was measured for its predictive accuracy.

Module 3: Machine Learning Lab I

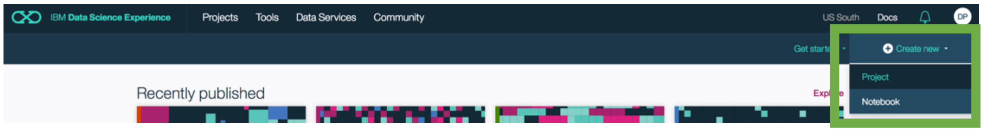
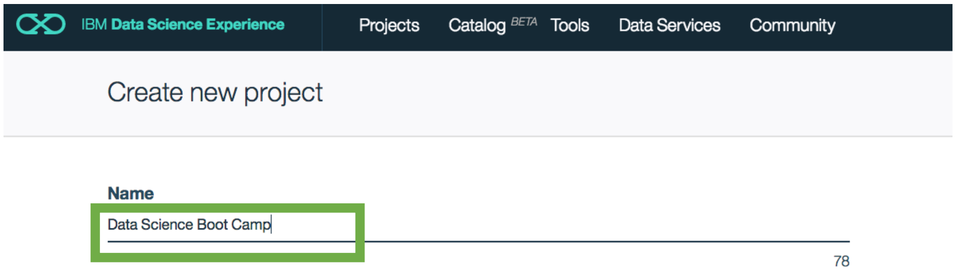
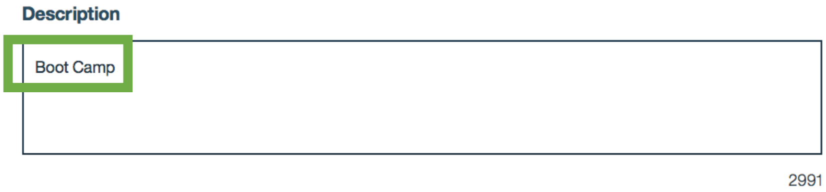
Purpose:	This lab introduces the tools used for a Machine Learning project written in R. After completing the lab, you should be able to: • Pull data from a GitHub repository into a Jupyter notebook • Perform an exploratory analysis of a dataset in IBM's Data Science Experience (DSX) • Create training and testing datasets

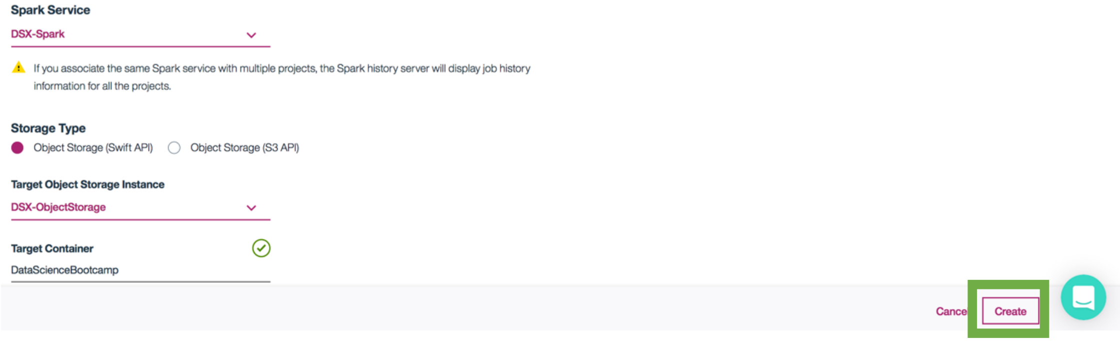
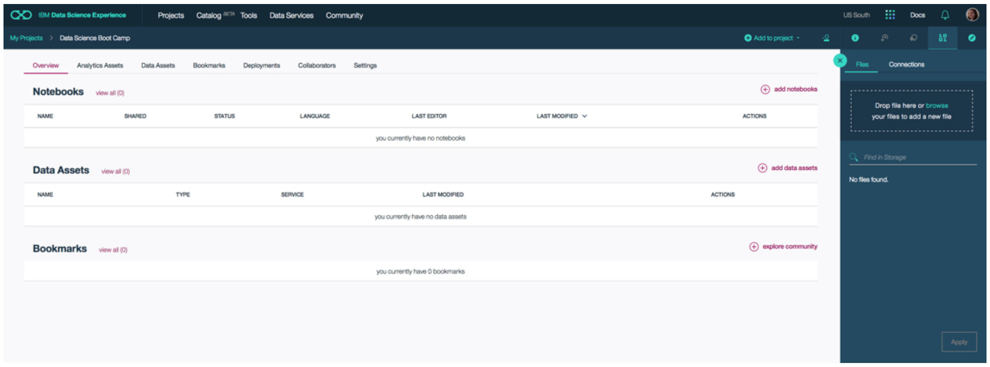
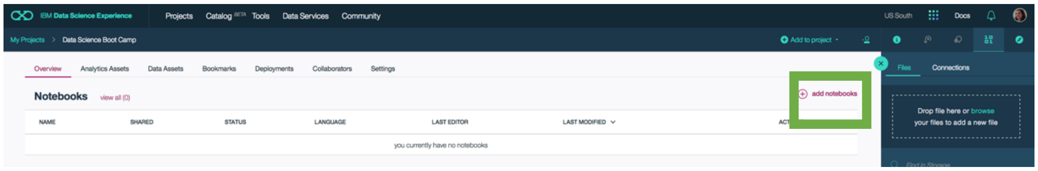
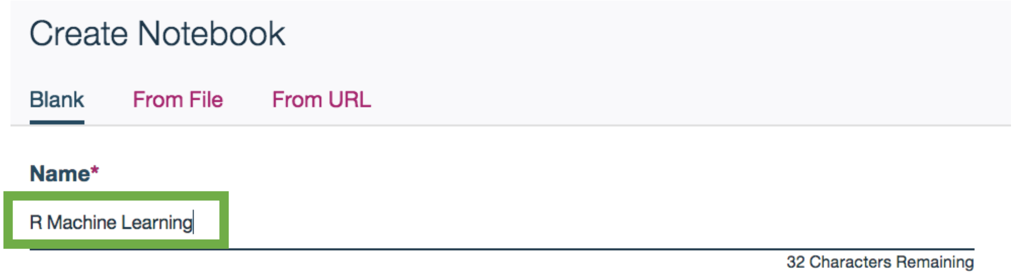
Tasks:	Tasks you will complete in this lab exercise include: • Install and load R libraries • Exploration and Analysis ○ Analyze the data using visualizations ○ Test the data for correlations ○ Create training and testing datasets

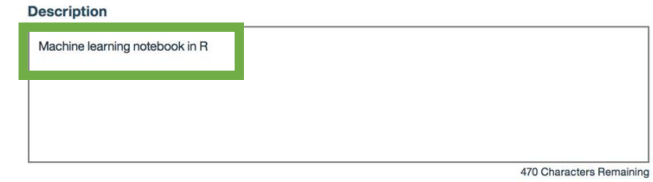
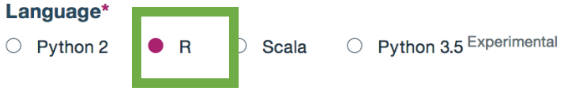
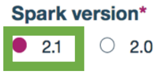
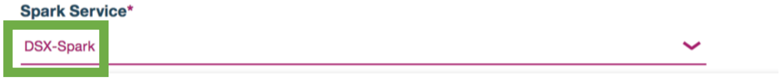
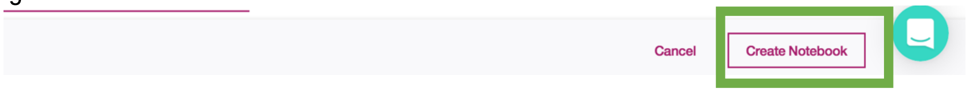
Module 3: Lab Workflow Overview

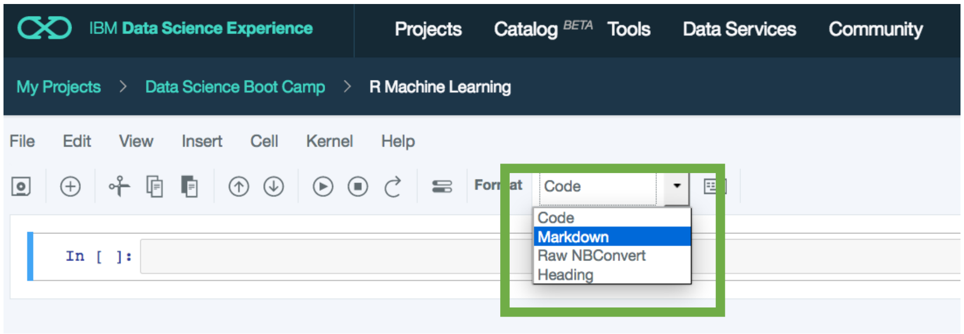
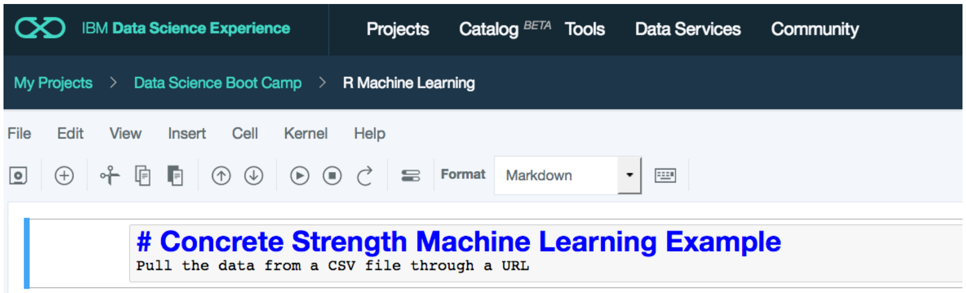
1. Open DSX and Create a new project
2. Create a notebook
3. Add a markdown title
4. Install necessary libraries
5. Load libraries
6. Pull data from GitHub
7. Set the seed
8. Check the data was loaded and the size
9. Examine the descriptive statistics
10. Observe the distribution of each variable
11. Determine the correlation value
12. Create training and testing datasets

Module 3: Lab Instructions

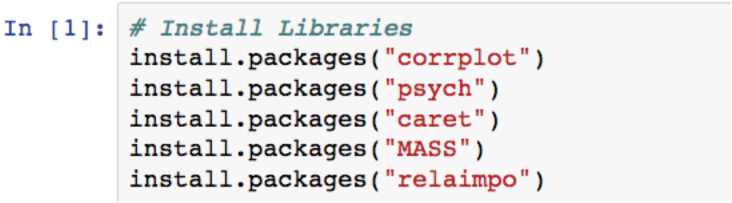
Step	Action
1	**Open DSX and Create a new project** a. Navigate to https://datascience.ibm.com/ b. Login to DSX c. On the top right side, click **Create New and select project**  IBM Data Science Experience Projects Tools Data Services Community US South Docs DP Get starte Create new Project Notebook Recently published Expl e d. Type the Project Name **Data Science Boot Camp**  IBM Data Science Experience Projects Catalog BETA Tools Data Services Community Create new project Name Data Science Boot Camp 78 e. Type the Description **Boot Camp** Description Boot Camp 2991 f. Ensure the defaults are selected as follows: Select your Spark Service **DSX-Spark** Select **Object Storage(Swift API)** Select Target Object Storage Instance **DSX-ObjectStorage** Default Target Container **DataScienceBootCamp**

Step	Action
	 **Spark Service** DSX-Spark ⚠ If you associate the same Spark service with multiple projects, the Spark history server will display job history information for all the projects. **Storage Type** ● Object Storage (Swift API) ○ Object Storage (S3 API) **Target Object Storage Instance** DSX-ObjectStorage **Target Container** ✓ DataScienceBootcamp g. Click **Create**  IBM Data Science Experience Projects Catalog BETA Tools Data Services Community US South Docs My Projects > Data Science Boot Camp Add to project Overview Analytics Assets Data Assets Bookmarks Deployments Collaborators Settings Files Connections Drop file here or browse your files to add a new file Find in Storage No files found. Apply **Notebooks** view all (0) ⊕ add notebooks NAME SHARED STATUS LANGUAGE LAST EDITOR LAST MODIFIED ACTIONS you currently have no notebooks **Data Assets** view all (0) ⊕ add data assets NAME TYPE SERVICE LAST MODIFIED ACTIONS you currently have no data assets **Bookmarks** view all (0) ⊕ explore community you currently have 0 bookmarks
2	**Create a notebook** a. Click **Add Notebooks**  IBM Data Science Experience Projects Catalog BETA Tools Data Services Community US South Docs My Projects > Data Science Boot Camp Add to project Overview Analytics Assets Data Assets Bookmarks Deployments Collaborators Settings Files Connections Drop file here or browse your files to add a new file Find in Storage **Notebooks** view all (0) ⊕ add notebooks NAME SHARED STATUS LANGUAGE LAST EDITOR LAST MODIFIED ACT you currently have no notebooks b. Type Notebook Name **R Machine Learning**  Create Notebook Blank From File From URL **Name*** R Machine Learning 32 Characters Remaining

Step	Action
	c. Type the Description **Machine Learning notebook in R** Description Machine learning notebook in R 470 Characters Remaining d. Select **R** for the language Language* ○ Python 2 ● R ○ Scala ○ Python 3.5 Experimental e. Select **2.1** for the Spark version Spark version* ● 2.1 ○ 2.0 ○ 1.6 f. Select the Spark Service **DSX-Spark** Spark Service* DSX-Spark g. Click Create Notebook 
3	**Add a markdown title** The behavior of a cell is determined by a cell's type. The different types of cells include: *Code*: Where you can edit and write new code. *Markdown:* Where you can document the computational process. You can input headings to structure your notebook hierarchically. *Raw NBConvert cells:* Where you can write output directly or put code that you don't want to run. Raw cells are not evaluated by the notebook. For the purpose of this lab, a heading will be added but all further notes will be inline with the code by using #. An example of using Markdown will follow.

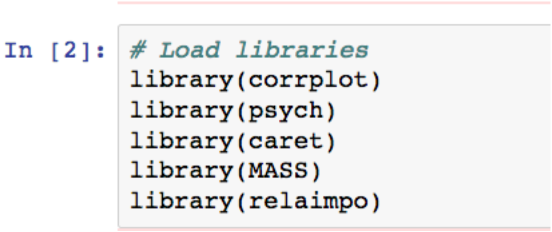
Step	Action
	a. Select the format to be **Markdown**  IBM Data Science Experience Projects Catalog BETA Tools Data Services Community My Projects > Data Science Boot Camp > R Machine Learning File Edit View Insert Cell Kernel Help Format Code Code Markdown Raw NBConvert Heading In [ ]: b. Type in the cell: # Concrete Strength Machine Learning Example Pull the data from a CSV file through a URL c. Click **run** (play button) or you can use the shortcut **“shift + enter**” to execute the cell.  IBM Data Science Experience Projects Catalog BETA Tools Data Services Community My Projects > Data Science Boot Camp > R Machine Learning File Edit View Insert Cell Kernel Help Format Markdown **# Concrete Strength Machine Learning Example** `Pull the data from a CSV file through a URL`
4	<u>**Install necessary libraries**</u> Many R functions come in packages, which are free libraries of code written by R's active user community. There are thousands of helpful R packages but this lab will only require the following: *Corrplot:* graphical display of a correlation matrix, confidence interval *Psych:* basic descriptive statistics useful for psychometrics *Caret:* set of functions that streamline the process for creating predictive models *MASS:* functions and datasets to support Venables and Ripley, ``Modern Applied Statistics with S''

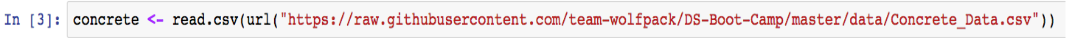
Step	Action

	Relaimpo: provides several metrics for assessing relative importance in linear models where they can be printed, plotted and bootstrapped
	a. Enter the code:
	# Install Libraries
	install.packages("corrplot")
	install.packages("psych")
	install.packages("caret")
	install.packages(“MASS”)
	install.packages(“relaimpo”)
	b. **Run** cell
	
	In [1]: # Install Libraries
	install.packages("corrplot")
	install.packages("psych")
	install.packages("caret")
	install.packages("MASS")
	install.packages("relaimpo")
	Note: Installing the libraries may take some time. Once installed, a red box will appear with an installation confirmation. This is normal and informational only. A similar red box will appear in the next step as well and is normal when loading libraries.

5	**<u>Load libraries</u>**
	Loading libraries gives you access to the functions that they contain. By using libraries, programmers can focus on the task at hand and not worry about developing functions that the user community has already developed.
	a. Enter the code:
	# Load libraries
	library(corrplot)
	library(psych)
	library(caret)
	library(MASS)
	library(relaimpo)
	Note: A red box will appear. This is normal and informational only.

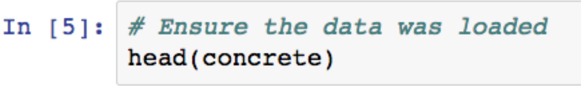
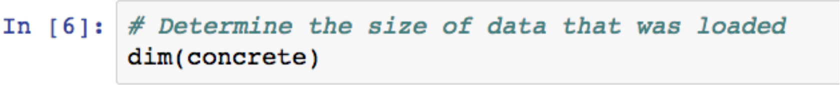
Step	Action

	b. **Run** cell
	
	`In [2]: # Load libraries`
	`library(corrplot)`
	`library(psych)`
	`library(caret)`
	`library(MASS)`
	`library(relaimpo)`

6	**<u>Pull data from GitHub</u>** Data can be brought into DSX in multiple ways. For this lab, you will pull a data file from GitHub related to concrete strength. The data will be stored as a data frame named concrete. A data frame is an in-memory storage format that is representative of the csv data, and accessed via a variable name. a. Enter the code: concrete <- read.csv(url("https://raw.githubusercontent.com/team-wolfpack/DS-Boot-Camp/master/data/Concrete_Data.csv")) b. **Run** cell  `In [3]: concrete <- read.csv(url("https://raw.githubusercontent.com/team-wolfpack/DS-Boot-Camp/master/data/Concrete_Data.csv"))`

7	**<u>Set the seed</u>**
	Generally, in statistics, samples are chosen at random. A random number generator is used to select the samples and is based off of a seed value. The seed is explicitly set so results are reproducible. To ensure everyone retrieves the same results in this lab, the seed value was randomly chosen as 3482.
	a. Enter the code
	# Set seed to ensure reproducibility
	set.seed(3482)
	b. **Run** cell
	
	`In [4]: # Set seed to ensure reproducibility`
	`set.seed(3482)`

Step	Action

8	**Check the data was loaded and the size**
	To check that data is present, the head command is used to retrieve the first few rows of data from the data frame specified.
	a. Enter the code
	# Ensure the data was loaded
	head(concrete)
	Note: You can access documentation pages for R functions, datasets and other objects directly by entering the command **?function**. Example: **?head** If you want more information on the commands used in this lab, you should access them to better understand the code being entered.
	b. **Run** cell
	
	In [5]: # Ensure the data was loaded
	head(concrete)
	c. For the observed greatest strength, what are the values for cement_amt___________ and water_amt___________?
	d. Enter the code
	# Determine the size of data that was loaded
	dim(concrete)
	e. **Run** cell
	
	In [6]: # Determine the size of data that was loaded
	dim(concrete)
	f. How many rows and columns are in the dataset?
	Rows:________
	Columns: ________
	Hint: You can use the help operator to learn about the dim() function output in order to answer this question. See Note in Step 8a.

Step	Action

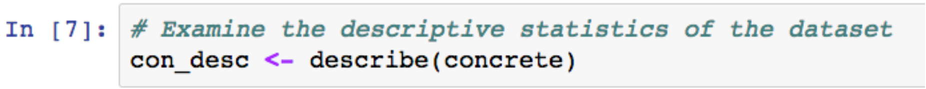
9 **Examine the descriptive statistics**

Descriptive statistics are used to describe or summarize features in a collection of data. These are broken down into some well-known components such as mean, median and standard deviation.

a. Enter the code

```
# Examine the descriptive statistics of the dataset
con_desc <- describe(concrete)
```

b. **Run** cell



```
In [7]: # Examine the descriptive statistics of the dataset
        con_desc <- describe(concrete)
```

c. Enter the code

```
con_desc
```

d. **Run** cell

```
In [8]: con_desc
```

e. Write down the median for plast_amt: _________

10 **Observe the distribution of each variable**

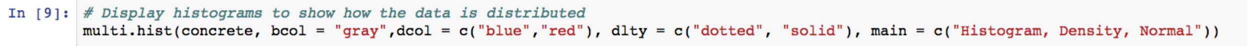
Histograms allow for the distribution of values to be seen very quickly. Values are counted in bins that consist of ranges of values. The taller the bar, the larger the count of values for the range is.

a. Enter the code

```
# Display histograms to show how the data is distributed
multi.hist(concrete, bcol = "gray",dcol = c("blue","red"), dlty = c("dotted", "solid"), main = c("Histogram, Density, Normal"))
```

b. **Run** cell

Step	Action



```
In [9]: # Display histograms to show how the data is distributed
        multi.hist(concrete, bcol = "gray",dcol = c("blue","red"), dlty = c("dotted", "solid"), main = c("Histogram, Density, Normal"))
```

11

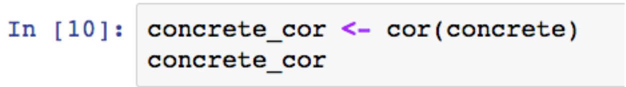
Determine the correlation value

Correlation shows how two variables relate to each other. The value of the correlation represents a percentage of change that is related between the variables. If the correlation is positive, both variables move in the same direction. Negative correlation means the variables move in opposite directions. A correlation of 1 indicates both variables move the same amount together. If correlation is zero, there is no relationship in how the variables move together.

a. Enter the code

```
concrete_cor <- cor(concrete)
concrete_cor
```

b. **Run** cell



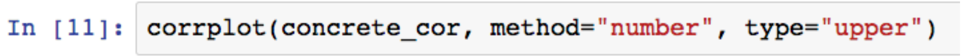
```
In [10]: concrete_cor <- cor(concrete)
         concrete_cor
```

c. Write down the correlation for water_amt and plast_amt: ____________

d. Enter code

```
corrplot(concrete_cor, method="number", type="upper")
```

e. **Run** cell



```
In [11]: corrplot(concrete_cor, method="number", type="upper")
```

12

Create training and testing datasets

In this section, the model will have the ability to adaptively resample the tuning parameter in order to concentrate on values that will provide the optimal settings.

First, the data will be split. 80% to train the model and 20% to test it.

“One of the first decisions to make when modeling is to decide which samples will be used to evaluate performance. Ideally, the model should be evaluated on samples that were not used to build or fine-tune the model, so that they provide an unbiased sense of model effectiveness. When a large amount of data is at hand, a set of samples can be set aside to evaluate the final model. The “training” data set is the general term for

Step	Action

the samples used to create the model, while the “test” or “validation” data set is used to qualify performance.” [(1)]

In most cases, the training and test samples are desired to be as homogenous as possible. Random sampling methods can be used to create similar datasets.

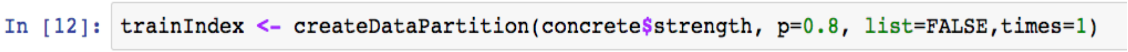
Example:

“Assume that we need to estimate average number of votes for each candidate in an election. Assume that country has 3 towns: Town A has 1 million factory workers; Town B has 2 million office workers and Town C has 3 million retirees. We can choose to get a random sample of size 60 over entire population but there is some chance that the random sample turns out to be not well balanced across these towns and hence is biased causing a significant error in estimation. Instead if we choose to take a random sample of 10, 20 and 30 from Town A, B and C respectively then we can produce a smaller error in estimation for the same total size of sample.” [(2)]

a. Enter the code

```
trainIndex <- createDataPartition(concrete$strength, p=0.8,
list=FALSE,times=1)
```

b. **Run** cell



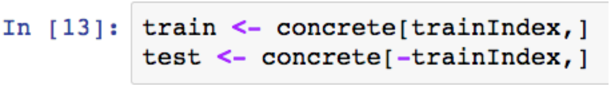
```
In [12]: trainIndex <- createDataPartition(concrete$strength, p=0.8, list=FALSE,times=1)
```

The split data will be labeled, train and test.

c. Enter the code

```
train <- concrete[trainIndex,]
test <- concrete[-trainIndex,]
```

d. **Run** cell



```
In [13]: train <- concrete[trainIndex,]
         test <- concrete[-trainIndex,]
```

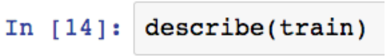
The data in train can be viewed.

Step	Action

e. Enter the code

```
describe(train)
```

f. **Run** cell



```
In [14]: describe(train)
```

g. Write down the median for plast_amt ________

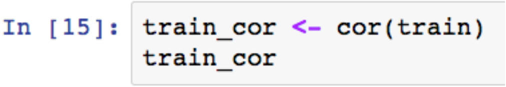
Examine Training Dataset

Note: We want to ensure that the splitting of the data did not result in different profiles for the training and testing data. The function used does its best to ensure the resulting datasets have a similar profile, but the best practice is to check.

h. Enter the code

```
train_cor <- cor(train)
train_cor
```

i. **Run** cell



```
In [15]: train_cor <- cor(train)
         train_cor
```

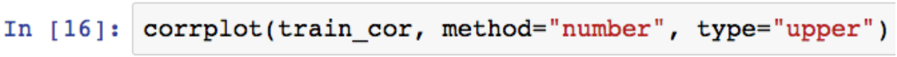
j. Write down the correlation for water_amt and plast_amt: ________

Plot the correlations for train.

k. Enter the code

```
corrplot(train_cor, method="number", type="upper")
```

l. **Run** cell



```
In [16]: corrplot(train_cor, method="number", type="upper")
```

Examine Testing Dataset

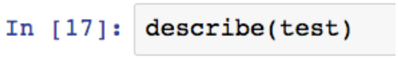
The data in test can be viewed.

Step	Action

m. Enter the code

describe(test)

n. **Run** cell



```
In [17]: describe(test)
```

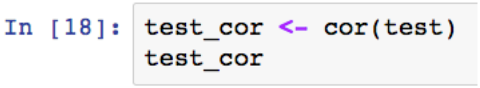
o. Write down the median for plast_amt: ________

Find the correlations for test.

p. Enter the code

test_cor <- cor(test)
test_cor

q. **Run** cell



```
In [18]: test_cor <- cor(test)
         test_cor
```

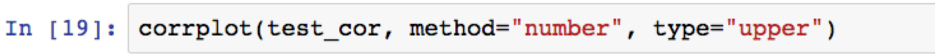
r. Write down the correlation for water_amt and plast_amt: ________

Plot the correlations for test.

s. Enter the code

corrplot(test_cor, method="number", type="upper")

t. **Run** cell



```
In [19]: corrplot(test_cor, method="number", type="upper")
```

Note: At this time, the training and test data is ready for Module 4: Approaches to Machine Learning.

Module 3: Lab Summary

Module 3 Lab started with IBM Data Science Experience (DSX). A project was created followed by an R notebook. A brief introduction to cell types was provided and a title was added to the notebook using Markdown. The required libraries, also called packages, were loaded. Then, a dataset was loaded from a CSV file stored on GitHub.

To ensure reproducibility, a seed value was set at beginning of the lab. A seed value is used for random number generation and utilized for the random selection process when the training and testing datasets are created.

The first few rows of the dataset were visually inspected along with the size. Descriptive statistics were leveraged to better understand the data. To describe the data, visualizations were implemented.

Variable relationships were measured to determine which were strongly correlated. Correlations range from 0 to 1, where 0 indicates there is no correlation and 1 means there is a strong correlation. Values between 0 and 1 mean there is some relationship between the two variables and high correlation generally starts at a value of 0.60.

The final step of this lab was to create the training and testing datasets which will be utilized in the Module 4 Lab.

References

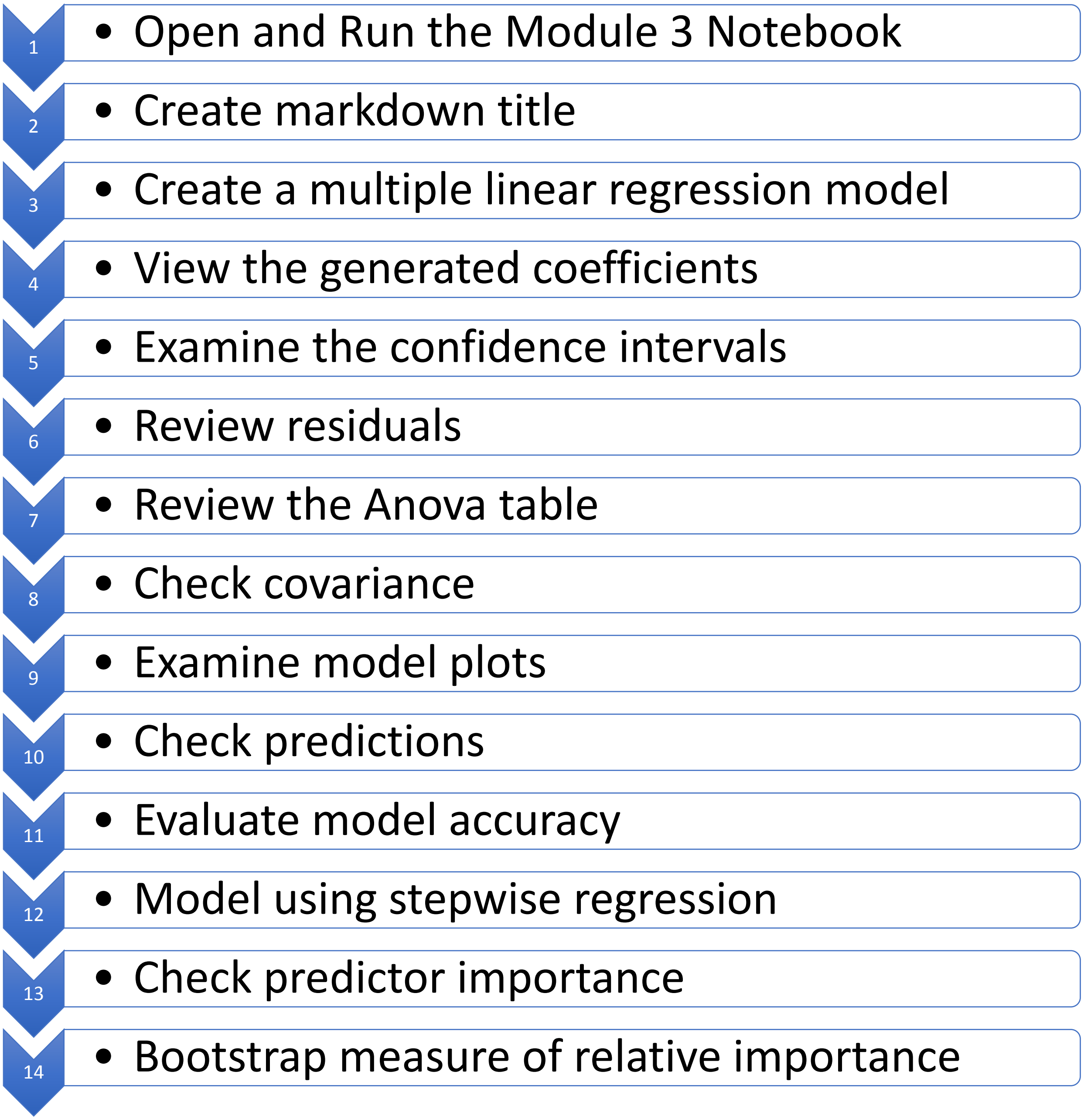
[1] Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (pp. 389-400). New York: Springer.
[2] Stratified Sampling. Wikipedia. https://en.wikipedia.org/wiki/Stratified_sampling

Module 4: Machine Learning Lab II

Purpose:	This lab is a continuation of the Module 3 Lab. After completing the lab, you should be able to: • Utilize a training dataset to train models • Interpret and test the model • Evaluate model accuracy • Explain model results • Choose the best model

Tasks:	Tasks you will complete in this lab exercise include: • Train multiple models • Test multiple models against a testing dataset • Compare the different models' performance

Module 4: Lab Workflow Overview



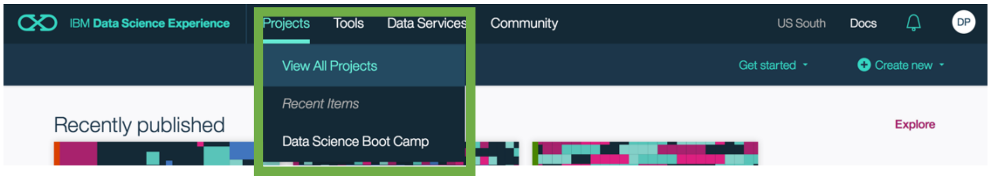
1. Open and Run the Module 3 Notebook
2. Create markdown title
3. Create a multiple linear regression model
4. View the generated coefficients
5. Examine the confidence intervals
6. Review residuals
7. Review the Anova table
8. Check covariance
9. Examine model plots
10. Check predictions
11. Evaluate model accuracy
12. Model using stepwise regression
13. Check predictor importance
14. Bootstrap measure of relative importance

Step	Action
1	**Open and Run the Module 3 Notebook**

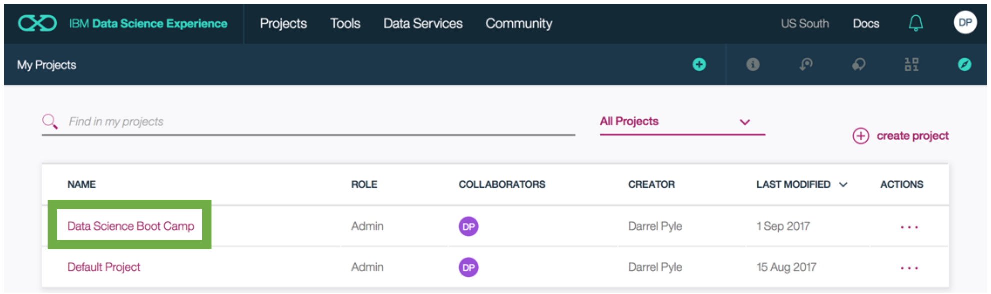
a. Navigate to https://datascience.ibm.com/

b. Login to DSX

c. In the top menu bar, click **Projects** and select **View All Projects**

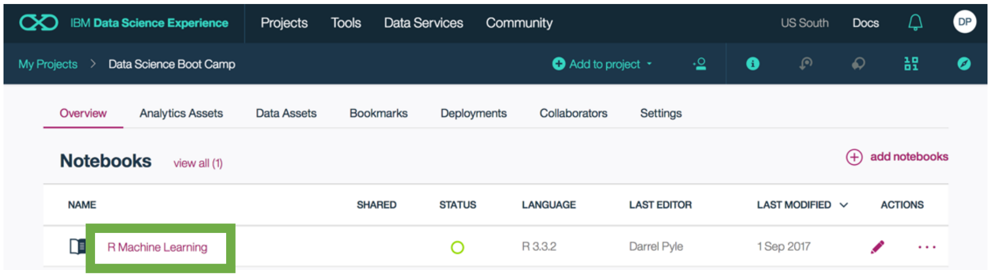


d. Click **Data Science Boot Camp**



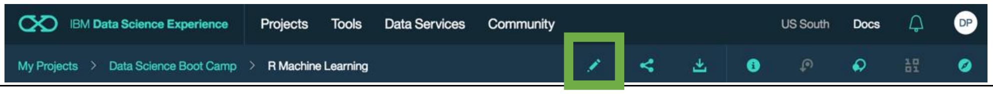
NAME	ROLE	COLLABORATORS	CREATOR	LAST MODIFIED	ACTIONS
Data Science Boot Camp	Admin	DP	Darrel Pyle	1 Sep 2017	...
Default Project	Admin	DP	Darrel Pyle	15 Aug 2017	...

e. Click **R Machine Learning** to open the notebook that was created in Module 3



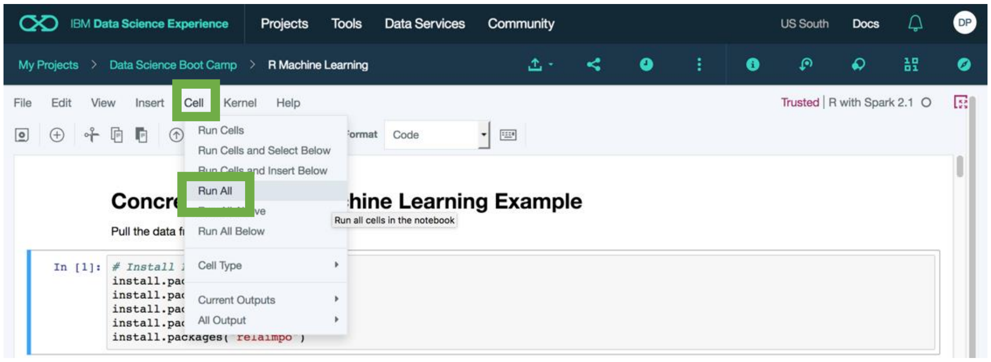
NAME	SHARED	STATUS	LANGUAGE	LAST EDITOR	LAST MODIFIED	ACTIONS
R Machine Learning		○	R 3.3.2	Darrel Pyle	1 Sep 2017	...

f. Click the **edit** (pencil) icon in the toolbar

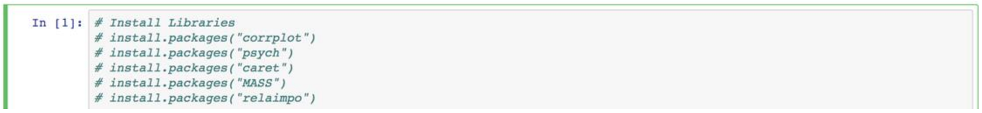

IBM Data Science Experience
Projects
Tools
Data Services
Community
US South
Docs
DP
My Projects > Data Science Boot Camp > R Machine Learning

Step	Action

g. From the menu bar, select **Cell** and click **Run All**



By running all of the cells, the referenced libraries are installed and loaded into memory to be available for usage. If the libraries have already been installed, typically the rows with the install.packages statements would be commented out by placing # in front of each line or the rows would be deleted. This would reduce the script execution time. Below is an example of commenting out the install.packages lines.



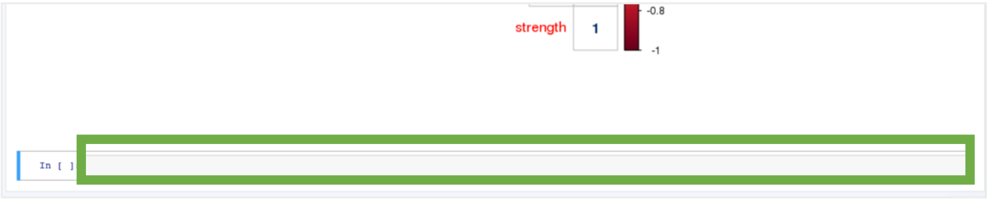
```
In [1]: # Install Libraries
        # install.packages("corrplot")
        # install.packages("psych")
        # install.packages("caret")
        # install.packages("MASS")
        # install.packages("relaimpo")
```

Data is also loaded into the data frames and are made available to the script. All the remaining commands are also executed. This places the notebook at the same point as it was when Module 3 Lab was completed and it is ready to have additional script added to it.

2 **<u>Create markdown title</u>**

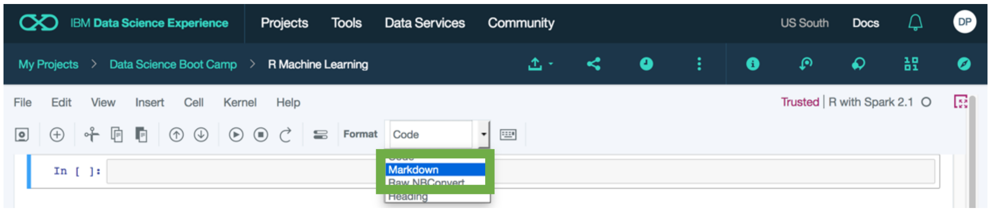
Markdown will be used to place a title to separate the Module 4 Lab code from the code of the previous exercise.

a. Scroll to the bottom of the notebook and click in the last empty cell



Step	Action

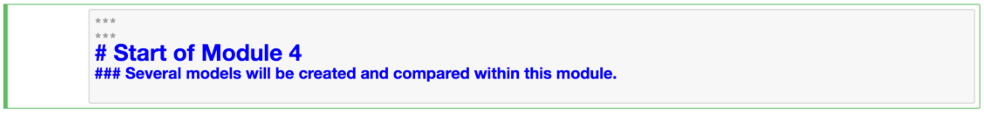
b. Change the cell to Markdown in the format dropdown in the toolbar



c. Type in the cell:

***
***
# Start of Module 4
### Several models will be created and compared within this module.

d. **Run** the cell



```
***
***
# Start of Module 4
### Several models will be created and compared within this module.
```

e. What do the "***" characters create? ______________________

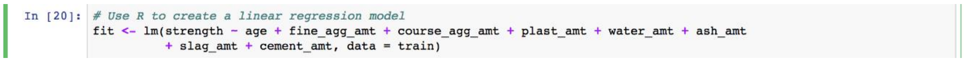
3 **Create a multiple linear regression model**

Multiple Linear Regression determines the relationship between a dependent (target) variable and the remaining independent (attribute) variables. The relationship of a linear regression equation assumes that each independent variable affects the dependent variable in a linear and additive manner.

a. Type in the cell:

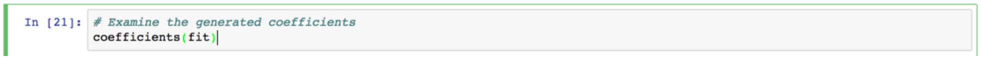
```
# Use R to create a linear regression model
fit <- lm(strength ~ age + fine_agg_amt + course_agg_amt +
plast_amt + water_amt + ash_amt + slag_amt + cement_amt, data =
train)
```

b. **Run** the cell



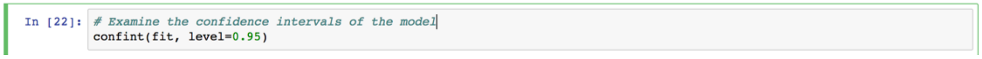
```
In [20]: # Use R to create a linear regression model
         fit <- lm(strength ~ age + fine_agg_amt + course_agg_amt + plast_amt + water_amt + ash_amt
                   + slag_amt + cement_amt, data = train)
```

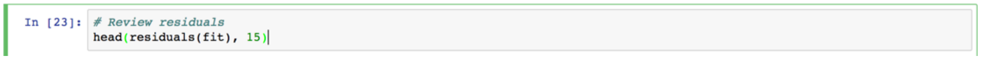
Step	Action
	Strength (located before ~) is the dependent variable. The remaining variables (located after ~) are the independent variables. The multiple linear regression model is stored as a variable named fit.

4	**View the generated coefficients**
	Coefficients indicate the amount of impact an independent variable has on the dependent variable, the larger the coefficient, the greater the influence an independent variable has on the dependent variable. The sign of the coefficient indicates if the relationship of the independent variable to the dependent variable is negative or positive. The intercept indicates a base starting point for the dependent variable when the independent variables are not present.
	a. Type in the cell:
	# Examine the generated coefficients
	coefficients(fit)
	b. **Run** the cell
	
	In [21]: # Examine the generated coefficients
	coefficients(fit)
	c. The coefficient of course_agg_amt is: _________
	d. Does course_agg_amt have a positive or negative relationship to strength? ___________________________
	e. Does this indicate that course_agg_amt's effect increases or decreases strength? ___________________________

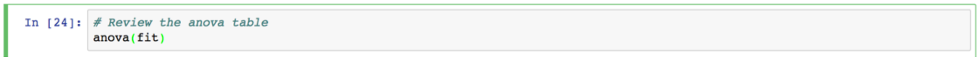
5	**Examine the confidence intervals**
	A confidence interval indicates the range of values that may contain the mean of the variable being looked at for a given probability (level). We will look at the confidence intervals that have a 95% probability (specified with “level=” parameter) of containing each variable.
	a. Enter the code:
	# Examine the confidence intervals of the model
	confint(fit, level=0.95)
	b. **Run** cell

Step	Action

	
	In [22]: # Examine the confidence intervals of the model
	confint(fit, level=0.95)
	c. The 95% confidence interval for ash_amt is:
	________________ to ________________
	The confidence interval means there is a 95% chance that the calculated interval contains the true mean (average) of ash_amt.

6	**Review residuals**
	In multiple linear regression, a residual is the vertical distance that a data point is from the line that is generated from the regression equation. The linear regression equation estimates the value of the dependent variable. The estimate is compared against the known data point to return the residual value. The closer the residual value is to zero, the closer the linear regression equation is to estimating the actual dependent value (i.e. the more accurate the model is).
	residual = known value – predicted value
	When the residual value is negative, it indicates the predicted value is too large which is known as an overestimate. Underestimated predicted values result in a positive residual value.
	For this example, the residual of each data point is the difference between the regression value of strength and the known strength value for each data point.
	a. Enter the code:
	# Review residuals
	head(residuals(fit), 15)
	b. **Run** cell
	
	In [23]: # Review residuals
	head(residuals(fit), 15)
	c. Which data point has the smallest residual? ________
	d. What is the residual value of the data point? ________
	e. Did the regression equation overestimate or underestimate the strength value? ________

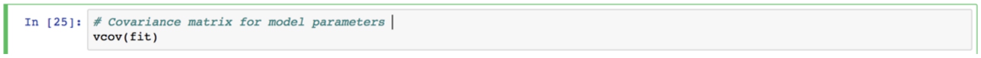
Step	Action

7	**Review the Anova table**
	An Anova table is another way to measure which variables have a significant impact on the dependent variable. To determine if a variable is significant, the p-value column is assessed and compared to the level of significance being used. Typically, a 5% level of significance is used which means the model explains 95% of the data. The p-value is compared to the level of significance (0.05) and if it is less than the level of significance, the variable is determined to have an impact on the dependent variable.
	The p-value for each variable is located in the “Pr(>F)” column corresponding to each variable in the Anova table.
	a. Enter the code
	`# Review the anova table`
	`anova(fit)`
	b. **Run** cell
	
	`In [24]: # Review the anova table`
	`anova(fit)`
	c. Which variable has the most significant impact (smallest p-value) on strength? _________________

8	**Check covariance**
	Covariance provides an indication of whether two variables increase or decrease together. If two variables increase or decrease together, the covariance will be positive. If one variable decreases while the other variable increases, then the covariance value will be negative. The magnitude of the covariance value indicates how far apart the values are from the mean; however, if the variables being compared are of different scales (such as feet compared to inches) the magnitude can be misleading. Typically, only the sign of covariance is used and the magnitude is ignored.
	a. Enter the code
	`# Covariance matrix for model parameters`
	`vcov(fit)`

Step | Action

b. **Run** cell



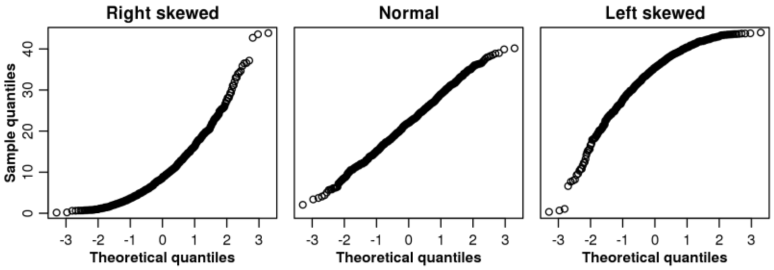
```
In [25]: # Covariance matrix for model parameters
         vcov(fit)
```

9

Examine model plots

The plots that will be examined are:

Residuals vs Fitted: Detect non-linearity, unequal error variance, and outliers.
Scale-location: Shows equally spread residuals along the ranges of predictors
Normal Q-Q: Check if residuals are normally distributed
Residuals vs Leverage: Helps find influential data points

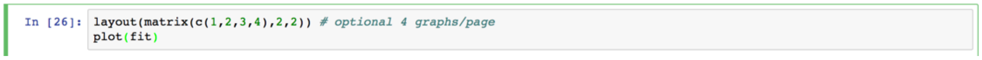


The figure above is a set of Q-Q plots with various data distributions. Right skewed data means the majority of data has small values with few large values. When the majority of data has large values with only a few small values, the data is left skewed. Normal data has most of its values surrounding the mean with decreasing amounts both smaller and larger than the mean. The smaller or larger the values, the fewer values there are, overall the data value count follow the bell curve. The Q-Q plotting the residual and quantity of the residuals. When a Q-Q plot is normal, illustrated by the middle graph above, it means the model predicts values higher and lower than the actual value with equal frequency. If the Q-Q plot is not normal, then the model is not accounting for everything that it needs to

a. Enter the code

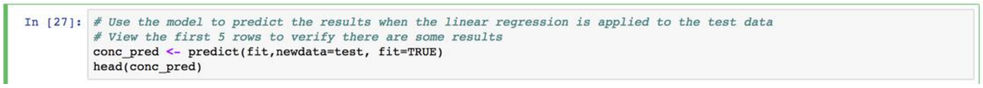
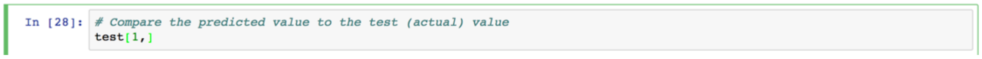
```
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(fit)
```

b. **Run** the cell

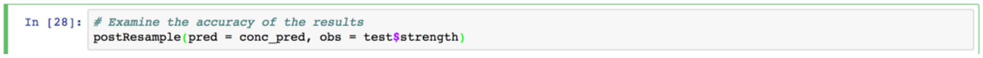


```
In [26]: layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
         plot(fit)
```

Step	Action
	c. The Normal Q-Q plot shows if residuals are normally distributed. This is indicated if the data points line up along the diagonal dotted line in the plot. Do the residuals appear to be normally distributed? ________

10	**<u>Check predictions</u>**
	The model can be applied to a testing dataset to calculate predicted values. The predicted values can then be compared against the actual values.
	a. Enter the code
	# Use the model to predict the results when the linear regression is applied to the test data
	# View the first 5 rows to verify there are some results
	conc_pred <- predict(fit,newdata=test, fit=TRUE)
	head(conc_pred)
	b. **Run** the cell
	
	In [27]: # Use the model to predict the results when the linear regression is applied to the test data
	# View the first 5 rows to verify there are some results
	conc_pred <- predict(fit,newdata=test, fit=TRUE)
	head(conc_pred)
	c. What is the value predicted for row ID 10? ________
	d. Enter the code
	# Compare the predicted value to the test (actual) value
	test[1,]
	e. **Run** the cell
	
	In [28]: # Compare the predicted value to the test (actual) value
	test[1,]
	f. What is the actual value of strength for row ID 10? ________
	g. What is the percent difference between the predicted and actual value?
	______________

Step	Action

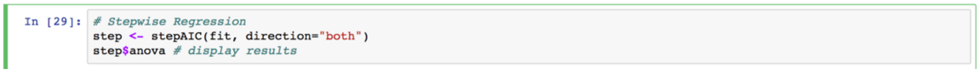
11	**Evaluate model accuracy**
	The value of R-Squared can be used to get a feel for the accuracy of predictions. The value ranges from 0% to 100% expressed as a decimal value. Generally, the higher the value the more accurate the model is.
	a. Enter the code
	`# Examine the accuracy of the results`
	`postResample(pred = conc_pred, obs = test$strength)`
	b. **Run** the cell
	
	`In [28]: # Examine the accuracy of the results`
	`postResample(pred = conc_pred, obs = test$strength)`
	c. What percentage of the data is explained by the model? _________

12	**Model using stepwise regression**
	Another modelling technique uses stepwise selection. Stepwise regression can function in one of three methods:
	Forward Selection: starts with no variables and adds in variables until model improvements cannot be made
	Backward Elimination: starts with all variables and removes variables until model improvements cannot be made
	Bidirectional Elimination: combines both forward and backward methods until model improvements cannot be made
	The overall goal of stepwise regression is to optimize which variables are included in a model. For the lab, bidirectional elimination will be used.
	a. Enter the code
	`# Stepwise Regression`
	`step <- stepAIC(fit, direction="both")`

Step	Action

step$anova # display results

b. **Run** the cell



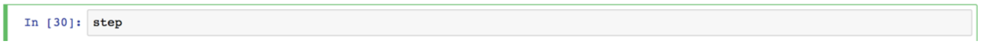
```
In [29]: # Stepwise Regression
         step <- stepAIC(fit, direction="both")
         step$anova # display results
```

Next, examine the stepwise model.

c. Enter the code

step

d. **Run** the cell



```
In [30]: step
```

e. The coefficients from the linear regression were:

(Intercept)	-6.69842457185358
age	0.111665479837888
fine_agg_amt	0.0150081212898075
course_agg_amt	0.0136261119756282
plast_amt	0.297816548355918
water_amt	-0.178034408123492
ash_amt	0.0775641746632639
slag_amt	0.0963375880785001
cement_amt	0.112662766607026

Does the stepwise regression model have different coefficients?

____

If the coefficients are the same, that can be an indication that all variables are adding value to the model and removing any particular variable will reduce the accuracy of the model.

13 **Check predictor importance**

Predictor importance can change based on when a variable is added to a model. Analyze how adding variables at different times affects the model. Four types of predictor importance will be examined:

LMG: utilizes R squared

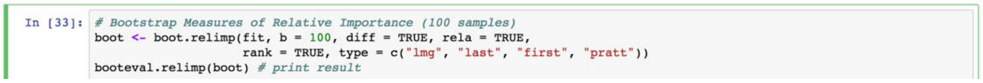
Step	Action

	Last: each variables contribution when included last
	First: each variables contribution when included first
	Pratt: utilizes the product of the standardized coefficient and the correlation
	
	a. Enter the code
	
	# Calculate Relative Importance for Each Predictor
	calc.relimp(fit,type=c("lmg","last","first","pratt"), rela=TRUE)
	
	b. Run the cell
	
	In [32]: # Calculate Relative Importance for Each Predictor
	calc.relimp(fit,type=c("lmg","last","first","pratt"), rela=TRUE)
	
	c. Which type of predictor importance has the variable with the largest relative importance metric? ________
	
	d. Which variable has the largest importance? ________
	
	e. What is the value of the importance? ________

14	**Bootstrap measure of relative importance**
	
	Bootstrapping is any test or metric that relies on random sampling with replacement. Random sampling with replacement means that after each sample is taken, the data point is placed back into the dataset and is available once again for the random sample process. Random sampling can also be performed without replacement. In this scenario, when a sample is taken, it is removed from the dataset and not available for further sampling. For this lab, the “with replacement” option is being used. Specifying a parameter for the number of bootstrap runs (b) to execute allows for greater validation and potentially more accuracy.
	
	a. Enter the code
	
	# Bootstrap Measures of Relative Importance (100 samples)
	boot <- boot.relimp(fit, b = 100, diff = TRUE, rela = TRUE, rank = TRUE, type = c("lmg", "last", "first", "pratt"))
	booteval.relimp(boot) # print result

Step	Action

b. **Run** the cell



```
In [33]: # Bootstrap Measures of Relative Importance (100 samples)
         boot <- boot.relimp(fit, b = 100, diff = TRUE, rela = TRUE,
                             rank = TRUE, type = c("lmg", "last", "first", "pratt"))
         booteval.relimp(boot) # print result
```

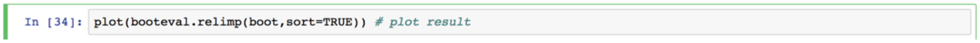
c. This produced similar output to the previous step to check the predictor relative importance. Is there anything that appears to be considerably different by using the bootstrap method? ________

To make it easier to identify the most important predictors, plots can be used.

d. Enter the code

```
plot(booteval.relimp(boot,sort=TRUE)) # plot result
```

e. **Run** the cell



```
In [34]: plot(booteval.relimp(boot,sort=TRUE)) # plot result
```

f. Which predictor is the most important for each predictor type?
LMG: ________
Last: ________
First: ________
Pratt: ________

Module 4: Lab Summary

At the end of Module 3’s Lab, training and testing datasets were created and used for this lab. Training data is used to create and train various machine learning models. Testing data is used to validate a trained model. It contains the actual data values for the dependent variable (strength) in addition to all of the independent variables.

The model was run against the independent variables (inputs) and calculated a value for the dependent variable (output). Once complete, the calculated value was compared against the actual value. The closer the calculated value was to the true value, the more accurate the model. These concepts were supported by many of the activities in the Module 3 Lab.

A multiple linear regression model was created. With this kind of model, the data typically appears to have a linear relationship. In this case, each independent variable was multiplied by a coefficient and all the values were added together to arrive at a value for the dependent variable.

For this model, here is the specific equation generated (shown with rounded coefficients):

strength = -6.6984 + 0.1117*age + 0.0150*fine_agg_amt + 0.0136*course_agg_amt
+ 0.2978*plast_amt - 0.1780*water_amt + 0.0776*ash_amt + 0.0963*slag_amt
+ 0.1127*cement_amt

Finally, model accuracy and performance were measured using confidence intervals, residuals, Anova, covariance, and bootstrapping.