BrainStation BC LLP
1-800-903-5159
vancouver@brainstation.io

Data Science Diploma Program

Student Transcript

A minimum passing grade of 70% is required for completion of the Data Science Diploma Program.

Units	Weight	Grade
Unit 1: Data Fundamentals	**15%**	90.0%
Unit 2: Analysis for Data Science	**20%**	94.7%
Unit 3: Machine Learning Techniques	**20%**	28.0%
Unit 4: Big Data Fundamentals	**15%**	
Unit 5: Professional Development	**30%**	
Total	100%	

Cumulative Grade: 69.2%

Student ID

Student ID:	169295
Alireza	Alizadegan

Data Science Diploma Program	
Start Date:	January 6, 2020
End Date:	March 27, 2020

Program Completion	
Status:	In Progress
Transcript Issued:	28-Feb-2020
Withdrawal Date (if Applicable)	

Attendance Grade

Below is a weekly breakdown of your attendance record. A minimum attendance grade of 90% is required for completion of the Data Science Diploma program. Please notify your TA of your absence with a reason ahead of time.

Week	Mon	Tue	Wed	Thur	Fri	Grade
1	7.0	7.0	7.0	7.0	7.0	100.0%
2	7.0	7.0	7.0	7.0	7.0	100.0%
3	7.0	7.0	7.0	7.0	7.0	100.0%
4	7.0	7.0	7.0	7.0	7.0	100.0%
5	7.0	7.0	7.0	7.0	7.0	100.0%
6	7.0	7.0	7.0	7.0	7.0	100.0%
7	0.0	7.0	7.0	7.0	7.0	80.0%
8	7.0	7.0	7.0	7.0	0.0	80.0%
9						0.0%
10						0.0%
11						0.0%
12						0.0%

Overall Attendance Grade: 63.3%

Grading System

The BrainStation grading system employs a numerical marking system. Below is a description of grade meanings.

Grade Meanings	Numerical Scale of Marks
Excellent	90-100%
Very Good	80%-90%
Good	70%-80%
Developing	60%-70%
Limited	0-60%

BrainStation BC LLP
1-800-903-5159
vancouver@brainstation.io

Unit 1: Data Fundamentals — 15%

Student ID

Student ID:	169295
Alireza	Alizadegan

Unit Project		**Performance Rating**						
		Exemplary	Very Good	Satisfactory	Developing	Limited	Incomplete	Days Late
Categories	Comprehension		x					
	Execution		x					
	Communication		x					

Grade: 90.0%

Additional Comments

Overall decent understanding and application of mysql-based data analysis. The communication is clear and to the point.
A few tips for future assignments:
a. Pay attention to query formatting. Writing clean and structured queries will make it easier for viewers to read and debug.
b. Add a "Drop if exist" query before create views or tables. It benefits future alterations of the table.
c. Avoid expressions like "Month 7" (July) in Part 1 Q3.1.
d. Practice more on the data visualization. For example, the stacked bar graph in Part 2 Q1 is difficult to compare the monthly variations in 2016.
e. Read the requirements carefully. The tableau workbook is not accessible as the table "avg_trips_per_day" should be named "worktable1".

BrainStation BC LLP
1-800-903-5159
vancouver@brainstation.io

Unit 2: Analysis for Data Science 20%

Student ID

Student ID:	169295
Alireza	Alizadegan

Deliverable 1	**40% of Unit Grade**	**Performance Rating**						
		Exemplary	Very Good	Satisfactory	Developing	Limited	Incomplete	Days Late
Categories	Comprehension	x						
	Execution		x					
	Communication	x						

Grade: 96.5%

Deliverable 2	**60% of Unit Grade**	**Performance Rating**						
		Exemplary	Very Good	Satisfactory	Developing	Limited	Incomplete	Days Late
Categories	Comprehension		x					
	Execution	x						
	Communication		x					

93.5%

Unit Grade

Unit Grade: 94.7%

Additional Comments

Stats and Politics Part 1
Great job Alireza! Your code was super easy to read and report spot on. A couple of things:
- A trick to remove the "Unnamed: 0" column when reading a csv in Pandas is to add index_col=0 at the end! Ex. df_vote = pd.read_csv('../data/votes.csv', index_col=0)
- When replacing null values, try and see how the values are distributed and based on that information, imputing the data. If it is normally distributed the mean will be fine, but if the values are skewed, then the median works better!
- There are a couple other duplicate columns that you had missed (Native American and Amerindian, Other and Other Races)
- For your bar graph of the counties, you can rotate your graph to have the states on the y-axis, this would make the graph much easier to read! You can also make your graph bigger using this line of code: plt.figure(figsize=(20,10)).

Stats and Politics Part 2
Demonstrate proficiency with for-loops. Excellent job on the evaluation and improvements on your model. Only thing to pay attention is that you need to document the threshold you choose for R-square.
Also, you need to explain why 12% prediction error is acceptable. For example, if your model is for diagnostic or disaster prediction, 12% error is quite huge.

BrainStation BC LLP
1-800-903-5159
vancouver@brainstation.io

Unit 3: Machine Learning Techniques 20%

Student ID

Student ID:	169295
Alireza	Alizadegan

Deliverable 1	**40% of Unit Grade**	**Performance Rating**						
		Exemplary	Very Good	Satisfactory	Developing	Limited	Incomplete	Days Late
Categories	Comprehension			x				
	Execution			x				
	Communication			x				

Grade: 70.0%

Deliverable 2	**60% of Unit Grade**	**Performance Rating**						
		Exemplary	Very Good	Satisfactory	Developing	Limited	Incomplete	Days Late
Categories	Comprehension							
	Execution							
	Communication							

Grade:

Unit Grade

Unit Grade: 28.0%

Additional Comments

Deliverable 1:
Good job Alireza! Your code was great and very easy to follow. There are a few things to note:
- Because of how your report was formatted, there are a number of questions that didn't have a clear answer, or wasn't answered at all (Q1c, 1d, 2a, 2b, 3e). For future assignments, please follow the order of the questions when formatting your assignment.
- Within the case_punc_stop_lemm function, you have nltk.download() in a couple lines. nltk.download() downloads the entire nltk library, and so every time the case_punc_stop_lemm is called, you download the entire nltk library two times! This caused a number of errors when I had opened the notebook and why the output for some of your cells are extremely large. For the future, try not to have any library downloading lines of code in your notebook.
- With changing the non-numeric columns to numeric, while there are a large number of unique values in Hotel_Address, Hotel_Name or Tags, we would have liked at least one to be One Hot Encoded, or some stronger reasoning as to why it can be dropped (ex: the size of the dataframe would be too large)