Declaration on Plagiarism

Name:	Group 10
Student Number:	19210200, 19210993, 19210675, 19210213, 19210749, 19211138, 19211275, 19210279
Programme:	MCM1
Module Code:	CA675
Assignment Title:	Cloud Technologies
Submission Date:	13 Dec 2019
Module Coordinator:	Dr Long Cheng

We declare that this material, which we now submit for assessment, is entirely our own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. We understand that plagiarism, collusion, and copying are grave and serious offenses in the university and accept the penalties that would be imposed should we engage in plagiarism, collusion or copying. We have read and understood the Assignment Regulations. We have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the source cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

We have read and understood the referencing guidelines found at
http://www.dcu.ie/info/regulations/plagiarism.shtml
https://www4.dcu.ie/students/az/plagiarism
and/or recommended in the assignment guidelines

Name: Satya Prakash (GC) **Date: 13-Dec-2019**

Introduction

'**Amovino'** is an online platform that helps to review several kinds of wine from all over the world. From the detailed description, price, type along with its origin to the customer's review are presented briefly on the website. Features of the website include the top 10 wines in the US and the top 20 most expensive wines in the world. We can also predict the price of a wine based on the points, country, and variety of the wine, which is achieved by using machine learning at the backend.

Overview of Dataset

The dataset has 150931 rows and 10 columns. The columns consist of the following fields:

1. ***Country***: This field contains the name of the country where the wine is from.
2. ***Description***: A detailed elucidation of each wine is provided in this field, including its ingredients, how old it is, where it's from and more.
3. ***Designation***: The name of the vineyard within the winery where the grapes that made the wine are from.
4. ***Points***: The number of points Wine Enthusiast rated the wine on a scale of 1-100 (reviews only for wines that score >=80)
5. ***Price***: The cost of a bottle of wine.
6. ***Province***: province or state that the wine is from.
7. ***Region_1***: The wine-growing area in a province or state.
8. ***Region_2***: Sometimes there are more specific regions specified within a wine-growing area, but this value can sometimes be blank.
9. ***Variety***: The type of grapes used to make wine.
10. ***Winery***: The name of the winery that made the wine.

Data Processing

Initially, there were numerous glitches in the data. Cleaning of the data involved, removal of null values, irrelevant commas, and blanks. Two columns were also removed including index and region as per the requirement of the project. Here are the steps for the cleaning of the data:

```
g2sagar771994@cluster-7ae4-m:~$ cat winemag-data_first150k.csv > wine.csv
```

i. cat winemag-data_first150k.csv > wine.csv

ii. hadoop fs -put wine.csv /CloudAssignment

```
csvFile  pig_1575124216980.log  pig_1575147527024.log  wine.csv              Wineoutput.csv

g2sagar771994@cluster-7ae4-m:~$ hadoop fs -put wine.csv /CloudAssignment
```

iii. Pig

iv. Assignment2data = LOAD '/CloudAssignment' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES_MULTILINE','NOCHANGE','SKIP_INPUT_HEADER') as (id :int,country :chararray, description:chararray, designation:chararray, points:int, price :int, province:chararray, region1:chararray, region2:chararray, variety:chararray, winery:chararray);

```
grunt> Assignment2data = LOAD '/CloudAssignment' USING org.apache.pig.piggybank.storage.CSVExcelS
LTILINE', 'NOCHANGE','SKIP_INPUT_HEADER') as (id: int,country :chararray,description :chararray,c
y,points:int,price :int,province:chararray,region1 :chararray,region2:chararray,variety:chararray
```

v. cleandata = FOREACH Assignment2data GENERATE country,REPLACE(REPLACE(description,',',''),',','') AS description,designation,points,price,province,region1,variety,winery;

```
grunt> cleandata = FOREACH Assignment2data GENERATE country,REPLACE(REPLACE(description,',',''),'
on,designation,points,price,province,region1,variety,winery;
```

vi. countdata = FOREACH (GROUP cleandata ALL) GENERATE COUNT(cleandata);

```
grunt> countdata = FOREACH (GROUP cleandata ALL) GENERATE COUNT(cleandata);
```

vii. dump counntdata

```
(150930)
```

viii. nonullfinal= Filter cleandata BY country != '';

```
grunt> nonullfinal= Filter cleandata BY country != '';
```

ix. countdata = FOREACH (GROUP nonullfinal ALL) GENERATE COUNT(nonullfinal);

x. dump countdata;

```
(150925)
grunt> countdata = FOREACH (GROUP nonullfinal ALL) GENERATE COUNT(nonullfinal);
```

xi. finalwine = FOREACH nonullfinal GENERATE FLATTEN((country,description,designation,points,price,province,region1,variety,winery));

```
grunt> finalwine = FOREACH nonullfinal GENERATE FLATTEN((country,description,designation,points,p
n1,variety,winery));
```

xii. limited = LIMIT finalwine 10;

xiii. dump limited;

```
(France,This is the top wine from La Bégude named after the highest point in the vineyard at 1200
ture density and considerable acidity that is still calming down. With 18 months in wood the wine
extra richness and concentration. Produced by the Tari family formerly of Château Giscours in Mar
ade for aging. Drink from 2020.,La Brûlade,95,66,Provence,Bandol,Provence red blend,Domaine de la
```

xiv. STORE finalwine INTO 'OUTPUTPIGSTORAGE.csv' USING PigStorage(',');

```
grunt> STORE finalwine INTO 'OUTPUTPIGSTORAGE.csv' USING PigStorage(',');
```

xv. hdfs dfs -ls

xvi. hdfs dfs -copyToLocal OUTPUTPIGSTORAGE.csv /home/g2sagar771994/

xvii. cd OUTPUTPIGSTORAGE.csv/

xviii. Downloading the file!!

```
part-m-00000    SUCCESS
```

```
g2sagar771994@cluster-7ae4-m:~$
g2sagar771994@cluster-7ae4-m:~$
g2sagar771994@cluster-7ae4-m:~$
```

File Transfer Cancel

part-m-00000 In progress

Working

1) **Front End:** The website is designed using only two technologies,

   a) HTML: Hypertext Markup language, is used to design the basic layout of the web pages
   b) CSS: Cascading Style Sheets is used for styling the layout of webpages, like text styles, visual effects, and background color.

2) **Backend**: Below queries were used to retrieve data from the back end to the front end.

1. CREATE TABLE wine(id float(10), country varchar(15),description varchar(999),points int(4),price int(5),variety varchar(20), winery varchar(100)); - This query is used to create a table in database. We loaded data from cloud bucket storage.
2. SELECT variety,country,description,price FROM wine1 WHERE country = 'US' AND variety!='' ORDER BY points DESC LIMIT 10; - This query retrieves top 10 wines of US based on their ratings.
3. SELECT variety, price, country, winery FROM wine1 WHERE country = 'US' ORDER BY price DESC LIMIT 20; - This query retrieves the top 20 most expensive wines within the US along with their description.
4. SELECT variety, points, country FROM wine1 WHERE variety LIKE '%<SEARCH_WINE_TEXT>%'; - This query works for the search operation of the website where users can search for specific wine and query returns the data like points, variety, and country for selected wine.

3) **Connectivity:** The data is initially uploaded on to the Google Cloud Platform. With the help of Cloud SQL, this data is then loaded into the tables.

   "*jdbc:mysql://google/<DATABASE_NAME>?cloudSqlInstance=<INSTANCE_CONNECTION_NAME>&socketFactory=com.google.cloud.sql.mysql.SocketFactory&useSSL=false&user=<MYSQL_USER_NAME>&password=<MYSQL_USER_PASSWORD>* " This JDBC string is used to connect the front end with Cloud SQL using JDBC.

4) **Website:** The website is created on the local machine and integrated with the help of GCP. This platform uses cloud SQL to save the data in the database. Cleaned file Part00000m.csv is uploaded in the database using a cloud storage bucket. Google app-engine project is created on eclipse and servlets are used to connect and display the required data and images on the front end.

Machine Learning

A. For machine learning, we are using the Linear Regression algorithm to predict the price of wine based on a variety of wine, country and rating points.

B. For linear regression, we need to formulate an equation that has at least two variables, a dependent and one or more independent. The first step is to decide the dependent and independent variables.
   - Dependent Variable: Price of wine
   - Independent Variable: Rating points, Country , Variety of wine
   - Regression equation: *F(Price) = a*(Points) + b*(Country) + c*(Variety) + d* (where, *a,b* and *c* are coefficients and *d* is constant)

C. Amongst our independent variables, only rating points were numerical value and the remaining two (country and variety) are categorical values. As linear regression requires numerical data only, we encoded the categorical values into numerical values using encoder function. After encoding categorical data, we got 48 countries and 625 distinct varieties of wines.

D. We used 70% of the data as a training dataset to train our model and 30% of the data was used as a test dataset to test the model.

E. Below is the graph for our machine learning model which shows the regression line for prices with respect to points.

F. This algorithm is implemented in Python and a file is created with the outcome which is then read into a servlet and displayed on the frontend. The linear regression graph for machine learning is created on python and displayed as a static image on the front end

ROLES AND RESPONSIBILITIES

Task Assigned	Task assigned	Student ID	Remarks
Satya Prakash	Working on the front end and designing of the website using various technologies like HTML and CSS and video creation required for the project.	**19210200**	Satisfactory
Anshika Sharma	Working on the UI to create the website based on the design and selection of data cleaning. Helping with report creation.	**19210993**	Satisfactory
Jasmeet Singh Narula	Look into the options to bring up the connectivity between the front-end website and the Cloud platform using GCP.	**19210675**	Satisfactory
Raj Singh	Cleaning of data in PIG to get the appropriate data and loading of the same in tables so as to apply the appropriate query. Worked on various other possible options to implement the connectivity. Midway Report Writing.	**19210213**	Satisfactory
Palash Dange	Worked on Eclipse dealing with the connectivity of Cloud SQL and Frontend. Created various servlets to show the data clearly on the frontend.	**19210749**	Satisfactory
Rohit S Toshniwal	Database design and structure, Cloud SQL connectivity with java using Eclipse platform, Machine Learning model implementation using Linear Regression, Testing and Training ML model, Midway Report Writing	**19211138**	Satisfactory
Saumya Gautam	Machine Learning model design, Encoding of categorical data for Linear regression, Database structure development, Documentation of Machine Learning model, Midway Report Writing	**19211275**	Satisfactory
Aniruddha Kulkarni	Worked on back end queries to retrieve data according to the questions proposed. I also worked on the documentation of the report.	**19210279**	Satisfactory

Links

Gitlab: https://gitlab.computing.dcu.ie/prakass2/ca675---assignment-2
Video: https://youtu.be/iB_3GeI1ABQ