Assignment

In this assignment there are two parts:

1. Data science (70%)
2. Regression models (30%)

Case Study:

You have been tasked by the emergency services of New York City to investigate the relationship between weather and the number of traffic collisions on a particular day. The emergency services would like you to build two regression models; a linear regression model and a Deep Neural Network regression model that can accurately predict the number of collisions on a particular day of the week which will allow them to optimise emergency response staff.

What do you need to do?

1. Create at least one set of data using BigQuery.
2. Host this/these csv data files in your GitHub repo for use with Colab
3. Create a Colab sheet with an R kernel. This is what you will submit for assignment 1. It should be noted that you will have to structure this like a report with working code. For example, a mixture of text and code with a good linear structure. Your report should contain the queries you used to create your data with explanation of your thinking i.e. context.
4. If you have carried out cleansing of data, you should export this data (or many pieces of data) and host them in GitHub for use in the python/tensorflow Colab sheet
5. Create a (or many) Colab sheets with a python kernel for creating your regression models. Again, this should be written like a report.

Some further points to think about. Collated data is at least what you need but try thinking about doing analysis on just the collision data (i.e. using long and lat and density diagrams). Think about how you can maybe use things like this in your collated data.

In case of issues with creating collated data, this simple collated weather and collision data file can be used for analysis and creating models: <https://raw.githubusercontent.com/1122131uhi/dataAnalytics/master/results-20200821-021149.csv>

Reports generally have the following structure:

1. Introduction
2. Methodology
3. Results
4. Conclusions
5. References

Assignment 1 has a word limit of 3500 words and assignment 2 has a word limit of 1500 words. This means that diagrams, tables and figures will play an important role in your submission. Please note, code does not count.

Submissions:

Part (i) – Save your Colab sheet to Github and submit the link with links to the data files you have used and created as well. Also, download as a .ipynb file and submit this in a zip file.

Part (ii) – Same as above but with the Colab sheet with your regression models in it.

These should be submitted to Brightspace.

Taught component, getting the data:

In this little section, I will help you find the correct data and show you how to collate the data.

Go to BigQuery:

A screenshot of a social media post

Description automatically generated

Click on the drop down arrow on the right of “ADD DATA”:

A screenshot of a cell phone

Description automatically generated

Then click on “Explore public datasets” and you should see the following:

A screenshot of a cell phone

Description automatically generated

Type in “nyc collisions” into the search box and press enter:

A screenshot of a cell phone

Description automatically generated

Click on the single result:

A screenshot of a cell phone

Description automatically generated

Something to note, the data provided is from 2012 to the present. So, we will need weather data over this period, which we have.

Then click “VIEW DATASET”:

A screenshot of a cell phone

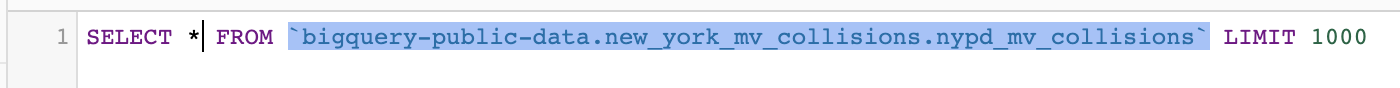
Description automatically generated

Click on the chevron and you can see the table:

A screenshot of a cell phone

Description automatically generated

Let’s have a look at the data, click on “QUERY TABLE” to generate a simple SELECT query:



Click on “Run”:

A screenshot of a cell phone

Description automatically generated

As you can see, we have timestamp. We will need this.

Now, just as in Tutorial 1, we need weather data. Remember, it’s in here:

bigquery-public-data > noaa\_gsod > gsod2009

Now, what we want to do, is go to our uhi-project

A screenshot of a cell phone

Description automatically generated

Then on the right, click on “CREATE DATASET”, then add an id in the window. I went for assignment\_guide. Obviously, you can pick what you want. After that click “Create dataset”.

A screenshot of a cell phone

Description automatically generated

Firstly, let’s create a view for every date available from 2012 to 2020. At the bottom (and throughout Tutorial 1, I show you how to do this, but let me show you again:

A screenshot of a social media post

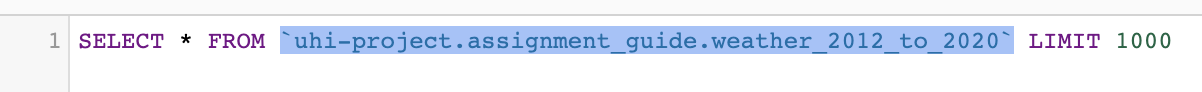
Description automatically generated

Then Run this:

A screenshot of a cell phone

Description automatically generated

Let’s take a look at the data:

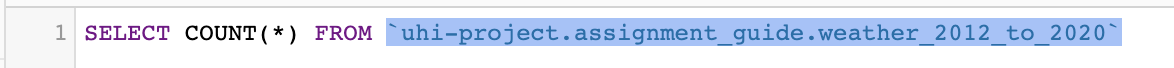


Which gives:

A picture containing computer

Description automatically generated

Looks fine, let’s check the number of records:



A close up of a logo

Description automatically generated

Hmm, is this correct….

2012 – 366

2013 – 365

2014 – 365

2015 – 365

2016 – 366

2017 – 365

2018 – 365

2019 – 365

2020 – 199 (last time I checked could be more here…)

I get 3121, which is about right considering it has been a while since I checked. Let’s just make sure:

SELECT \* FROM `bigquery-public-data.noaa\_gsod.gsod2020` WHERE stn='725060'

Thankfully it is 231 which gives 3153. Fantastic, our weather data looks perfect.

Next, let’s make a view that counts the number of collisions on a given day:

A screenshot of a social media post

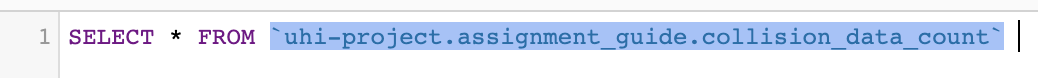
Description automatically generated

We get:

A screenshot of a cell phone

Description automatically generated

Now to check the data:



Looks good:

A picture containing bird

Description automatically generated

But you should note that there are only 2969 rows here which shows a discrepancy between our weather data. This is expected. Not all datasets are kept up to date as well as others.

Let’s add the day of the week column just like Tutorial 1:

A screenshot of a cell phone

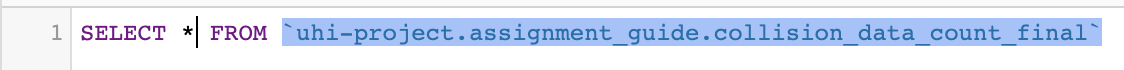
Description automatically generated

Run this:

A screenshot of a cell phone

Description automatically generated

And check the data in collision\_data\_count\_final:



This gives:

A screenshot of a cell phone

Description automatically generated

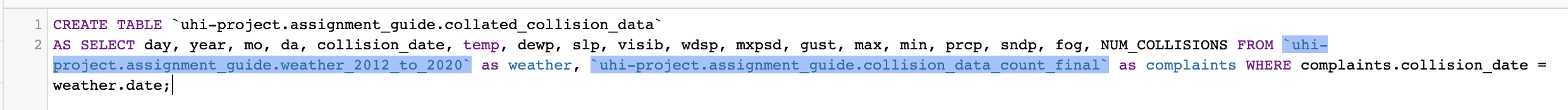
As you can see, the 4th of December 2019 is a Wednesday:

A screenshot of a cell phone

Description automatically generated

Data looks good so far.

Now, let’s finish things off by collating the data into a new table:



Run this and we get:

A screenshot of a cell phone

Description automatically generated

Check the data in collated\_collision\_data:



Giving:

A screenshot of a cell phone

Description automatically generated

Looks perfect and it has ignored the additional dates in the weather data.

Now we export it:

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Click on CSV (local file):

A close up of a logo

Description automatically generated

The file will be downloaded. Next upload to your GitHub:

A screenshot of a cell phone

Description automatically generated

Remember your commit message. Click “Commit changes”. Navigate on Github to your added file:

A screenshot of a social media post

Description automatically generated

Click on Raw:

A close up of a logo

Description automatically generated

And there we go. The link for this file, in case you can’t make your own variations, is at the top of this file.