Certificate in Introductory Data Analytics

[**https://github.com/katharine1600/Katharine\_ucd.git**](https://github.com/katharine1600/Katharine_ucd.git)

**Real World Scenario**

For my project I have used 3 data sets:

1. **USA Names**

This is a public dataset available on Kaggle.

<https://raw.githubusercontent.com/organisciak/names/master/data/us-names-by-decade.csv>

<https://www.kaggle.com/datagov/usa-names>

This dataset was created by the US Social Security Administration and contains the first names of babies from Social Security card applications for births that occurred in the United States. It is an interesting data set as it shows the rise and fall of names over the decades. From the list of names, we can observe links to celebrities, royalty, different ethnic backgrounds, and different religious affiliations. Using a list and a dictionary we will note that the number and range of names has increased over the decades.

1. **US Graduate schools’ admission**

This data set is also available on Kaggle.

[https://www.kaggle.com/tanmoyie/us-graduate-schools-admission-parameters](https://www.kaggle.com/tanmoyie/us-graduate-schools-admission-parameters/download)

The dataset is collected from  [www.kaggle.com/mohansacharya/](http://www.kaggle.com/mohansacharya/) graduate-admissions. It can be used to experiment with to show how an applicant’s different academic parameters determine their success in their application to a high-ranking university. I used a NumPy array to display the correlation between University Rating and the GRE score. (Graduate Record Examination – a standardised test in the US). I also explored visually the link between University ranking and CGPA scores which I have discussed below.

1. **“Penguins.csv”. I loaded it directly from the Seaborn library to my Jupyter notebook**

This dataset is available from the Seaborne library.

<https://github.com/mwaskom/seaborn-data>

This data set was selected as neither of the above contained any null values for my data cleansing.

**Importing Data**

The three data sets have been imported using **APIs** and I have provided the https above. The CSV files have been read into a panda’s data frames. I have also imported the following libraries.

* import pandas as pd. (for data frames)
* import seaborn as sns. (for graphs)
* import NumPy as np (for maths functions)
* import matplotlib. pyplot as plt (for graphs)
* from NumPy import median, mean (for maths calculations)

**Analyzing data**

**USA names**

I have used this data set to demonstrate the following:

* Inbuilt functions: *head (*) and *info ()* to get knowledge about the data set. There are four columns: gender, name, decade, and count.
* The shape function was used to count the total number of names; 108,384.
* *Isnull (). sum*. () was used to check the sum of any **missing values** in my data set. I observed that there were none.
* The *unique* () function was used on the decade column. My output was an array showing the 11 decades from 1910 to 2010 inclusive.
* I have defined a custom function “decadedict ()”. In this function I have used a **for loop** to loop through the dataset to count x – the number of names in a decade. Using **loc** on the data frame (us\_names\_df) I selected the column decade.
* I created a separate data frame (top\_us\_names\_df) where I grouped the top 5 names in each of the 11 decades using the **groupby** function. I created another data frame (bottom\_us names) of the bottom 5 names by decade also using the **groupby function**.
* We have no index (no primary key) in either data frame. As I wish to keep all the names, I used **concat to merge** the two data frames. A check of the new data frame (concat\_data) shows that we have 110 rows and 4 columns.

**Python**

* In “decadedict ()” I created **a list** of the decades and an empty **dictionary within a custom defined function.** This will create a collection of key value pairs. The keys are the decades, and the values are the number of names in each decade.
* The **for loop** facilitates the checking of the entire data set and the dictionary enables the decades and the count of names to be stored. I then printed the dictionary to enable us to see the number of names in each decade.
* Using the dataset number 2 above; US graduate admissions I reviewed it and I calculated the **correlation coefficient using a numpy array.** I hadexpected a high correlation between University ratings and GRE scores. I have graphed this and discussed below.

**Visualize**

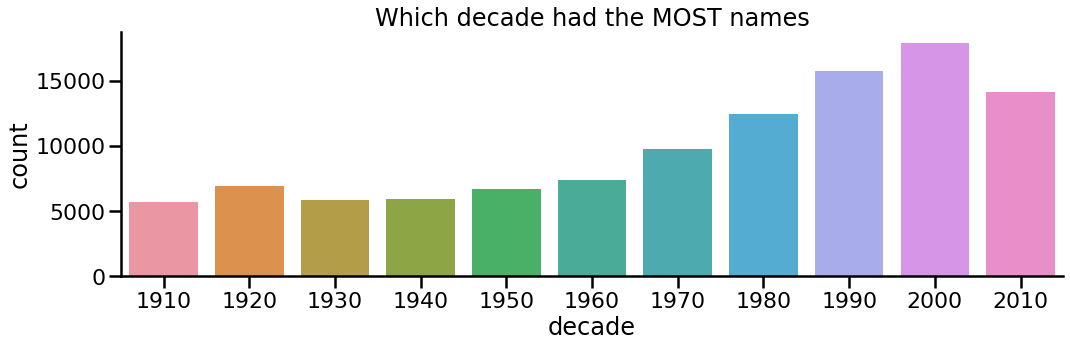
My first two visualisation are based on my first data set: USA Names. Graph number 3 is based on data set number 2; US graduate schools’ admission. This data set is a subset of the larger graduate admissions dataset referenced above. My final visualisation is based on the “Penguins” dataset from the Seaborn library. I have chosen plots from the seaborn (sns) library which is part of the matplotlib (plt) library for this exercise. Seaborn has a range of plots which are aesthetically pleasing with a colour palette suitable for my graphics.

**Graph 1**

**Background about the data set**

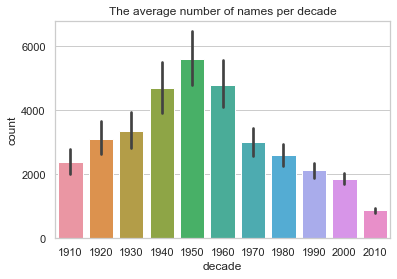
As observed above this data set spans 11 decades commencing in 1910 and up to and including 2019. It contains information about first names gleaned from social security card applications. This dataset has four variables: Gender, Name (First Name), Decade and Count. There are 108,384 observations. Two, Gender and Name are non-numeric qualitative variables while decade and count are quantitative variables.

To visualize the relationship between the two variables, first names (a qualitative variable) per decade and count (a quantitative variable) I chose a catplot. Names have been allocated to bins i.e., decades. The categorical estimate plots in the catplot function gives access to different plots that show relationships between numerical and categorical variables. Within the cat plot function, for the kind parameter I selected, count to count the number of names. The width of the bars is constant with the height reflecting the different number of names in each decade.



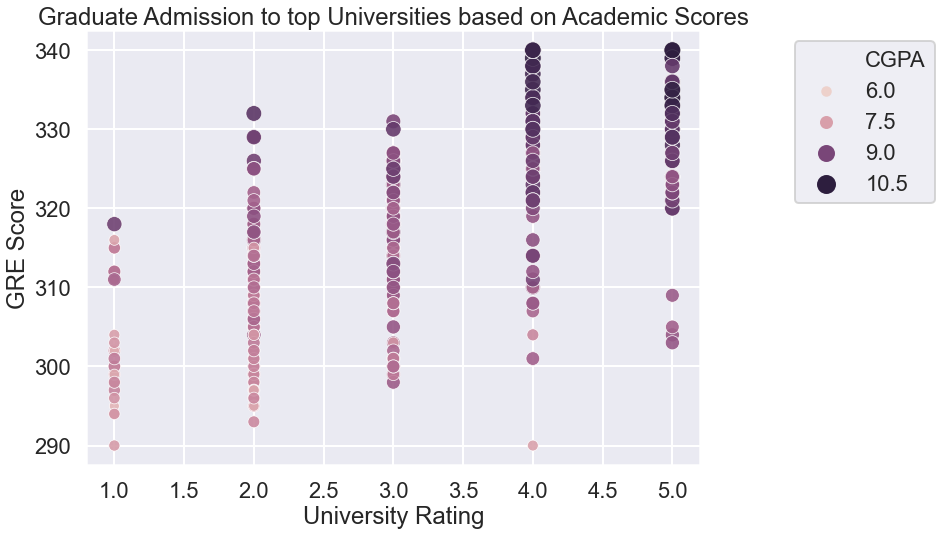
**Graph 2**

Graph2 is a bar plot showing the average number of names per decade. Using the estimator argument of the barplot () method in Seaborn you can alter how the data is aggregated. By default, each bin of a barplot displays the mean value of the variable (decade) the function has been used in combination with the list of names.



**Graph 3**

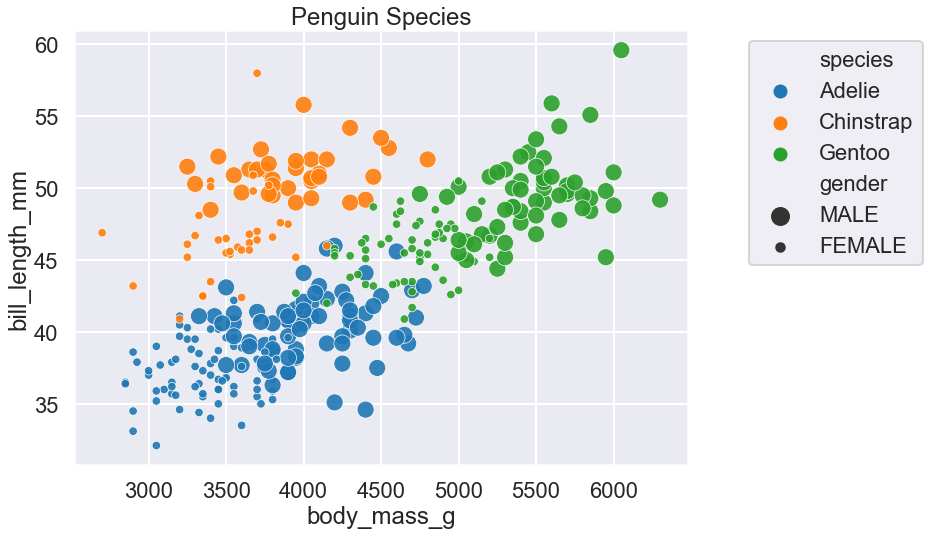
This graph looks at admissions to top US universities and the selection criteria. This data set is a sample of 400 records. Looking at an individual’s parameters including GRE scores (Grade Record Examination), SOP (Statement of Purpose) and CGPA (Cumulative grade Point Average). I sought a correlation between these parameters and admission to top universities.



**Graph 4**

This data set was selected as neither of the above contained any null values for my data cleansing.

I changed the column name from sex to gender and I checked for missing values. There were 11 rows with missing values and these rows were dropped from the data frame to give us 333 rows.



**Generate Valuable Insights**

**Graph** 1 – **Which decade had the most names.**

* The dataset has a qualitative variable on the X axis – names separated into bins(decades) and a quantitative variable on the Y axis the count of names.
* The total number of names of 108,384 have been split over the 11 decades. The number of names has increased decade on decade peaking with 17,862 names in 2000s up from 9,700 30 years earlier.
* Each decade has its own colour.
* The number of names was consistent circa 5,000 between 1910 and 1960 and increased significantly between 1970 and 2000.Note that whilst the number of names increased the US birth rate has been steadily declining over the last 30 years.
* The different names post 1960 would reflect the seismic post WWII demographic and societal changes reflecting a diverse multi-cultural society.

**Graph 2 – The average number of graphs per decade**

* Using the estimator argument, each decade is allocated a bin and each bin has the names for that decade. (like a histogram)
* The earlier decades with a tighter range of names have a higher average number of names reflecting a higher concentration of names.ie there were less first names in the first half of the 20th century, so the average number of each name was higher.eg the average number of Mary’s in 1950 was much higher than in 2010.
* In the second half of the century the number of names per decade increased but the average decreased as the range of names increased.
* The black line in the centre of the bin are error bars. These are graphical representations of the variability of data. They are used to indicate the error or uncertainty in a reported measurement.
* To reduce the error bar, we would need to change the confidence interval (ci) we would need to test for a lower ci.

**Graph 3 – Graduate Admissions to Top Universities based on Academic Scores.**

* With university ratings on the X axis ranked from 1 to 5, we note that as the rating of the university increases so do the GRE scores.
* The CGPA scores are reflected by the size and the colour of the circle. (see legend)
* The darker and larger the circle the higher CGPA score.
* This plot was selected to reflect the impact of both GRE and CGPA on your chances of gaining admission to a top university.
* The correlation coefficient for the top 40 scores is extremely high at .85 (see calculations) i.e., the top 40 scores have an extremely high probability of gaining admission to a high-ranking university but on the full data set the correlation drops to .66. Lower scores are predominately reflected in the lower ranking universities but there are some outliers at ranking 4 with a large spread at ranking 5. However, there is nobody with a CGPA of 10.5 at a university ranked below 4.

**Graph 4 – Penguin Species**

* My graph plots Penguins by bill length and body mass.
* Species are colour coded and analysed in the legend. Larger circles represent the male species.
* The three clusters of species are easily identifiable.
* There is little overlap between the species just some around the borders between the species.
* Outliers tend to move away from their cluster as opposed to overlapping another cluster.
* The ratio between male and female appears to be similar.

**Bibliography**

1.Waskom Michael seaborn. catplot <https://seaborn.pydata.org/generated/seaborn.catplot.html>

© Copyright 2012-2020, [Michael Waskom](https://www.cns.nyu.edu/~mwaskom). Created using [Sphinx](https://www.sphinx-doc.org/) 3.3.1

2.Data camp Modules 1-7

3. McQuaid David CCT College Dip 7 Notes

4.Stack Overflow (consulted regularly)

5.Visualization with Plotly.Express: Comprehensive guide | by Vaclav Dekanovsky | Towards Data Science  [https://towardsdatascience.com/visualization-with-plotly-express-comprehensive-guide-eb5ee4b50b57 consulted 16.04.2021-23.04.2021](https://towardsdatascience.com/visualization-with-plotly-express-comprehensive-guide-eb5ee4b50b57 consulted%2016.04.2021-23.04.2021)

6.IADT Data visualisation course notes 2019