

**Postgraduate Certificate in Cloud Native Computing**

**Postgraduate Certificate in Software Design with Artificial Intelligence**

**Applied Scripting Languages**

**Assignment 1**

Student ID: A00267948

Student Name: Daniel Foth

*Brief Description:*

*This assignment aims to design, implement and test a Python program to analyse and visualize a csv data set about information on bike sharing. In this document, the data, testing, visualisation and reflective learning will be described in detail.*

Contents

[Introduction 3](#_Toc22211610)

[Data 4](#_Toc22211611)

[Design 5](#_Toc22211612)

[Testing 6](#_Toc22211613)

[Analysis and Visualisation 7](#_Toc22211614)

[Conclusion 8](#_Toc22211615)

[Appendix 1: Reflective Learning Log 9](#_Toc22211616)

[Appendix 2: References 10](#_Toc22211617)

# Introduction

*This assignment aims to design, implement and test a Python program to analyse and visualize a csv data set about information on bike sharing. In the Data section below you can find information about the data, how its laid out and where it was sourced from.*

*The design for the program should be modular and allow for the importing of custom data sets to be processed via user input. It should be able to output graphs for linear regression and normal distribution. More information can be found in the design section below.*

*There are multiple ways to test code. As part of this assignment manual, unit and user testing shall be completed and described in the testing section below.*

*The analysis and visualisation of the data will help evaluate how successfully the data was analysed as part of the conclusion below.*

*Lastly a reflective log will be available detailing how the code was written and what issues were overcome.*

# Data

The data set was taken from <https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset>

This data set was composed by Hadi Fanaee-T and records the number of users interacting with a bike sharing system. It can give us an interesting view of how the bikes are utilised in different weather conditions. It also shows us how many registered/casual users use the services.

## Content

The data set includes 2 csv files. hours.csv and days.csv. hours.csv has 17 columns and 17389 entries. This will be the default dataset used for the analysis. days.csv has 16 columns and 731 entries.

## Columns

The following columns are present:

|  |  |
| --- | --- |
| Column | Description |
| instant | Index of the record |
| dteday | The date of the record |
| season | The season the record was taken in. (1=Springer, 2=Summer, 3=Fall, 4=Winter) |
| Yr | The year of the record. (0=2011, 1=2012) |
| mnth | The month of the record. (from 1 to 12) |
| hr | (not in days.csv) The hour of the record. (from 1 to 12) |
| holiday | Defines if a day is a holiday. (0=false, 1=true) |
| weekday | Defines if a day is a weekday. (0=false, 1=true) |
| workingday | Defines if a day is a working day meaning it’s a week day but not a holiday. (0=false, 1=true) |
| weathersit | This defines what the weather was like at the time of taking the record. (value from 1-4)   1. Clear, Few clouds, Partly cloudy, Partly cloudy 2. Mist & Cloudy, Mist & Broken clouds, Mist & Few clouds, Mist 3. Light Snow, Light Rain & Thunderstorm & Scattered clouds, Light Rain & Scattered clouds 4. Heavy Rain & Ice Pallets & Thunderstorm & Mist, Snow & Fog |
| temp | Normalized temperature in Celsius. (max is 41) |
| atemp | Normalized feeling temperature in Celsius. (max is 50) |
| hum | Normalized humidity. (max is 100) |
| windspeed | Normalized wind speed. (max is 67) |
| casual | Number of causal users per record |
| registered | Number of registered users per record |
| cnt | Total number of users per record |

## Formation

To compose this data Hadi Fanaee-T used 3 sources:

1. Original Source: <http://capitalbikeshare.com/system-data>
2. Weather Information: <http://www.freemeteo.com>
3. Holiday Schedule: <http://dchr.dc.gov/page/holiday-schedule>

## Why

This data set was chosen as it was interesting. There could be some interesting correlation between weather, workdays, holidays and the number of users of the Bike services. Maybe graphing this data in different ways could lead to some interesting results.

# Design

The project was designed with ease of development in mind. Therefore, git was used for version control and code was made modular when repetition occurred. The code and design developed over time in an agile like way as it was hard to define exactly what the application should be capable of from the very start.

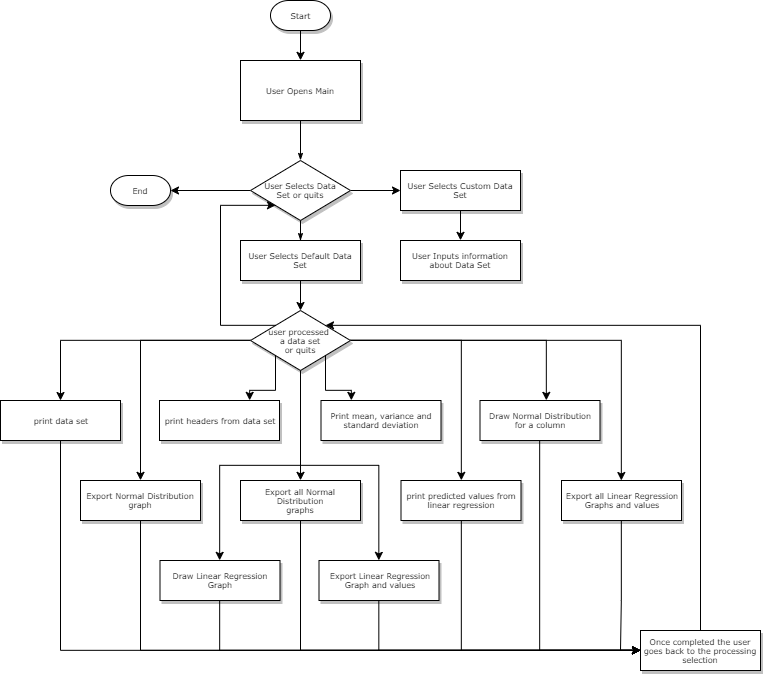
## Git

The GitHub repo can be found here: <https://github.com/DanielsHappyWorks/ASL-CA1-Data-Science-from-Scratch>. It shows how the design has been build incrementally with improvement over time.

## Program Flow

The python program can load in the provided data set or any other csv data file. When loading a custom file, it will ask the user where the file is, what are the column types and what is the delimiter. In both cases, when the file is loaded, the user can select how to process the file. He can print/export data based on normal distribution or linear regression using integer/floating-point columns in the dataset. Exporting data can be done for all columns at once for just one.

End to End program flow:

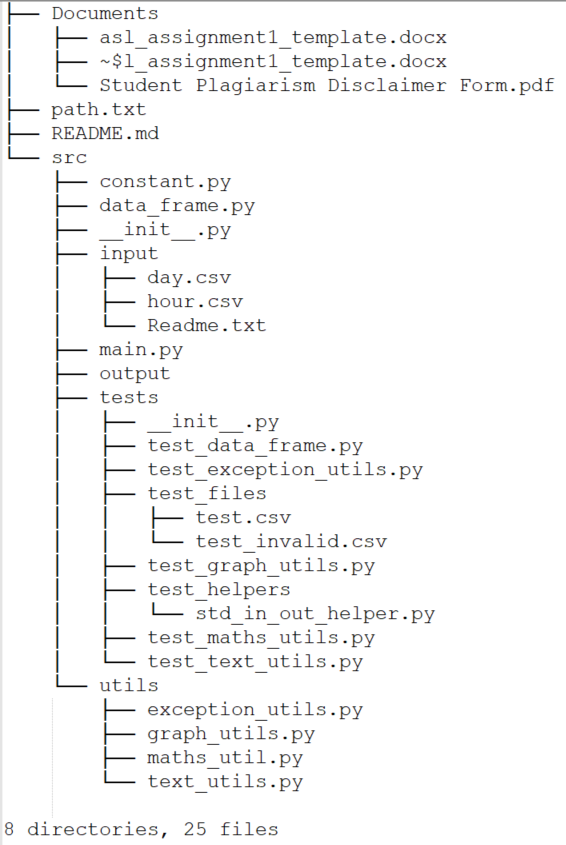


## Modularity

Keeping modular design in mind, classes were used to split up the code where applicable. The project consists of 3 main components:

1. Main Class – responsible for collecting user input and triggering processing on the DataFrame class.
2. DataFrame Class – responsible for processing of the data set with the help of the Utility Classes
3. Utility Classes – responsible for keeping the code maintainable and modular. This includes any maths, input and graphing utilities that are used in multiple places.

The current file structure for the design is laid out as below:



* Any documentation is present in the Documents directory or root of the project.
* All of the code is in the src directory and the main() function can be found in main.py.
* The files used for processing are stored in src/input and any output is put in src/output. These paths are accessed relatively so the project can be moved to different file systems.
* The tests are stored under src/tests with any helper functions in src/tests/test\_helpers.

# Testing

Three kinds of testing techniques were performed to test the python program. These include Manual testing, Unit testing and User testing.

## Manual Testing

The code was manually tested on a system with Python 3.7.4 [MSC v.1915 64 bit (AMD64)] on windows 32 mainly within the PyCharm Editor. Manual testing was done for each independent change. When the tiny independent changes made up a new feature the whole application was end to end tested with different flows in mind. The testing was done with the main data set and two test.csv files in the src/test directory.

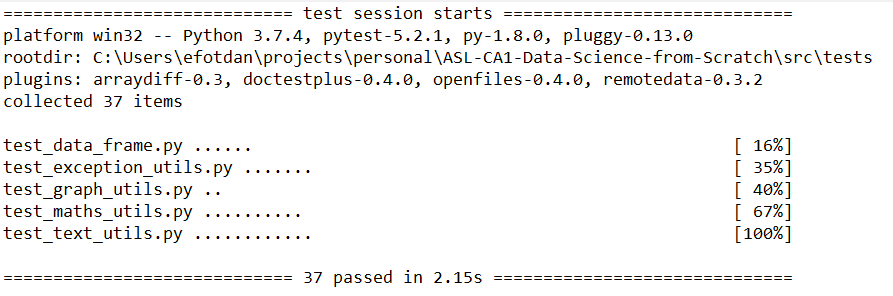
## Unit Testing

Due to time constraints and how the module was laid out the tests were mostly written after the coding for the different components was complete, but bugs were ironed out from the overall design using these. Thanks to unit tests some flows that weren’t tested often, one which broke over time, are now tested every push using automation. This made development more efficient as not everything needs to be tested now.

These unit tests can be triggered using pytest. They must be triggered manually from the src/tests directory as they are dependant on relative paths to get the test.scv files.

There are a lot of end to end flows relying on user input which are hard to unit test which lowers the overall test coverage.

Here is the current breakdown from pytest:



## User Testing

There were two user testing sessions to see how the application preformed. Both usability and weird inputs were tested by the users.

Session 1 Feedback (Around Nov 5th):

* More input validation needed – added extra validation so the application is more resilient to errors input
* Graph colours clashing and hard to look at – changed to similar cooler colours so they don’t clash
* Graphs were not clearing correctly – now clear graphs on reload
* Data did not load correctly on reload – now clear pyplot context before new graph is rendered
* Linear Regression doesn’t suit most of the default data set – this is very valid so extra analysis was taken and a Distribution graph was implemented for the next test

Session 2 Feedback (Around Nov 23rd):

* When selecting string column, the error prompt said to only select an integer – changed prompt to ask for an integer/float column
* What directory is the files exported to? –
* Single export seems to output to the wrong date –
* All linear regression export files should be consistent, start or end with lr –
* When reloading a custom data set the previous data types are used –
* When asking for data types the last one always ends up on 2 lines -

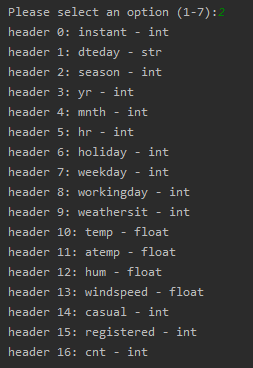
# Analysis and Visualisation

There are three types of data provided by the program. The information about the data set, normal distribution and linear regression outputs. The information about the data set is printed to the screen. Normal distribution and linear regression provide both graphical and text output. They can also be exported to files.

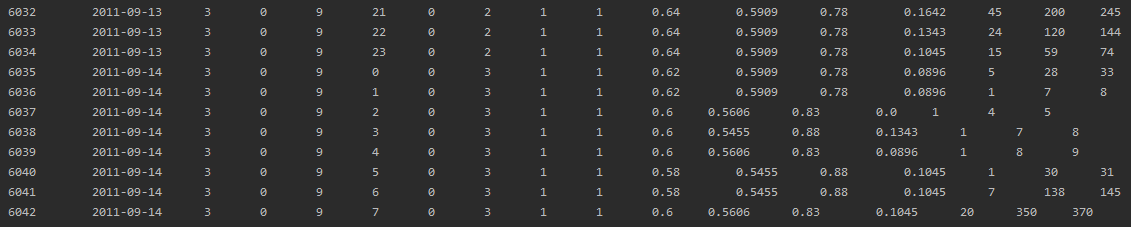
## Data Set Information

The whole data set can be printed to screen as a table separated using tabs. The headings can also be listed with the data types assigned to them.

Printed Headers:



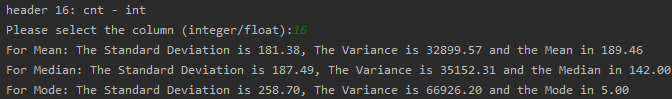
Printed Data:



## Normal Distribution.

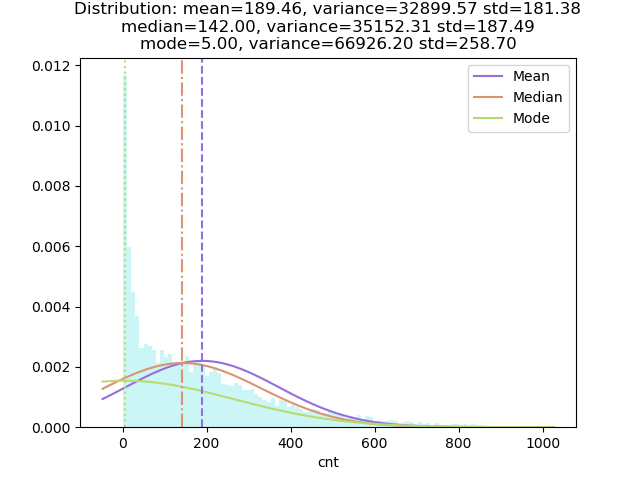
The data printable to screen based on this includes the mean, median and mode calculations for the selected column. For each of these the variance and standard deviation is also printed.

Sample print for total number of users from data set per record:



The graphs also display the previously mentioned values as text along with a histogram chart that has the distribution for each average overlaid above it.

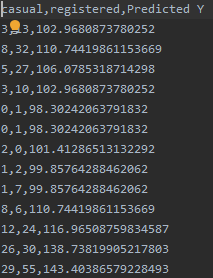
Sample graph for number of users (total) from data set per record:



## Linear Regression

For linear regression the predicted y values and the linear function to calculate them can be printed on screen. When exporting, these values will be outputted to a csv and txt file respectively.

Csv Output:

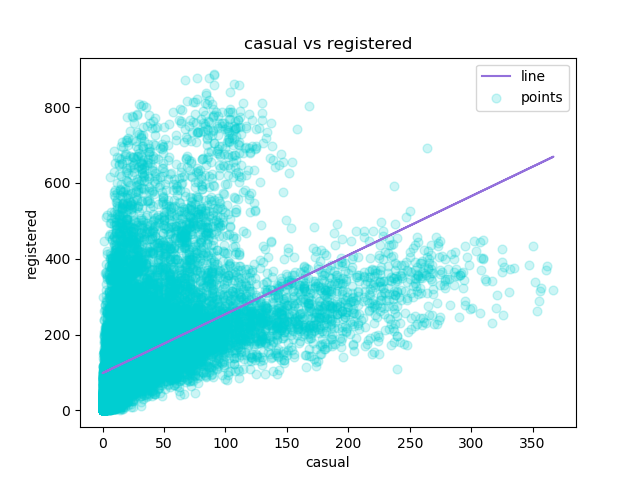


Formula Output:



The graphs for linear regression consist of two components. A scatter plot, which plots the x and y values against each other and a line which is plotted by using the x values and predicted y values using linear regression.

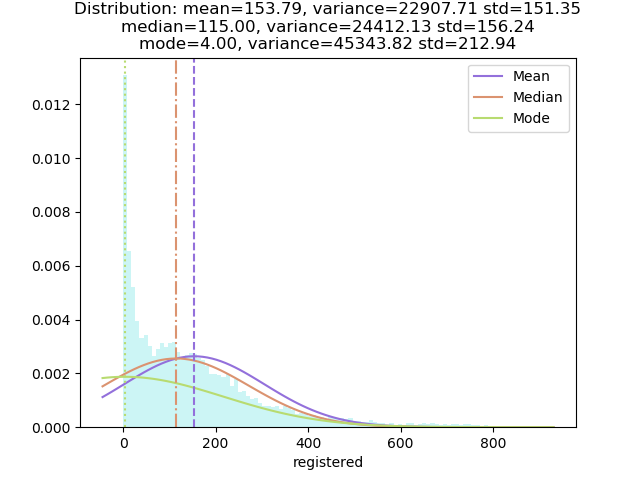
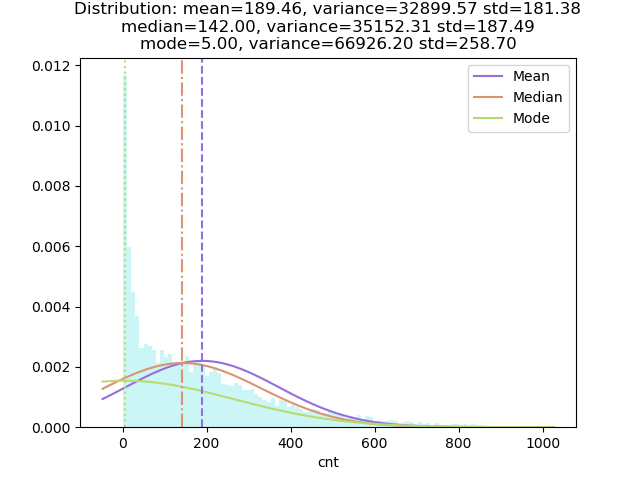
Plot of Casual vs Registered bike users:

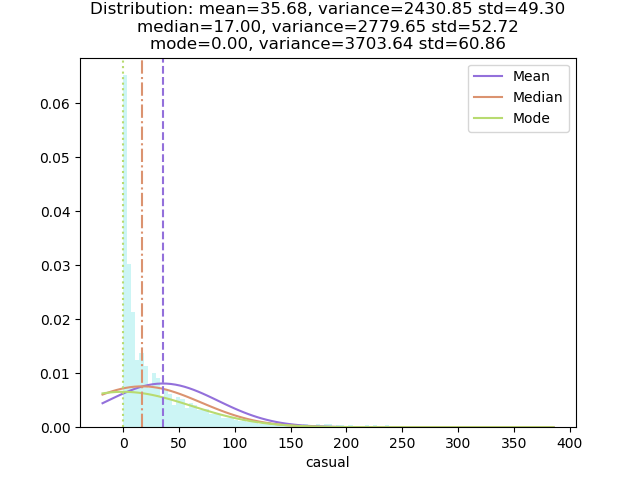


# Conclusion

## Normal Distribution.

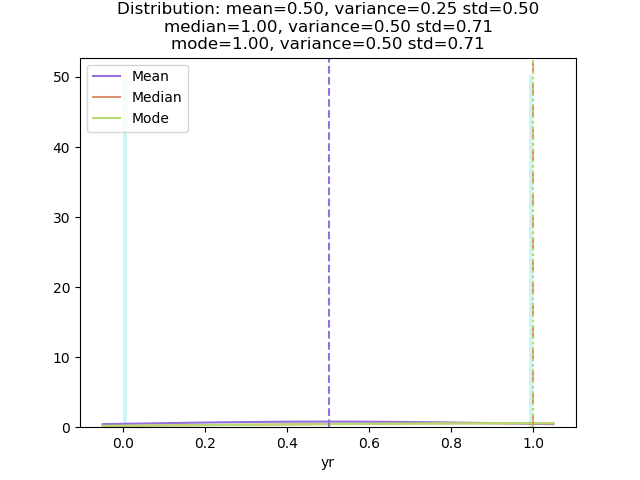
Sample graph for number of users (total, registered, casual) from data set per record:





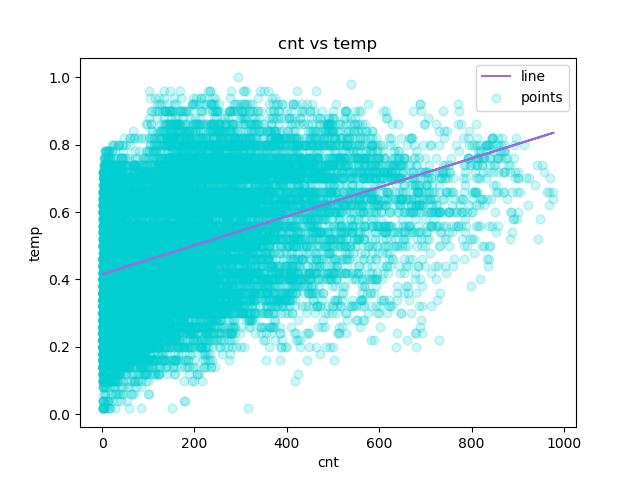
From samples like the above we can see that more registered users use the service then casual ones which gives us quick and interesting insights into the dataset and how the columns are distributed. This can be very useful when we quickly need to analyse a column.

On the other hand, we can end up with graphs like below which don’t really tell us anything at all about the dataset. This is because the program is generic enough to be able to plot any integer/float values. In these cases, we need human expertise to figure out if the output is useful or not.

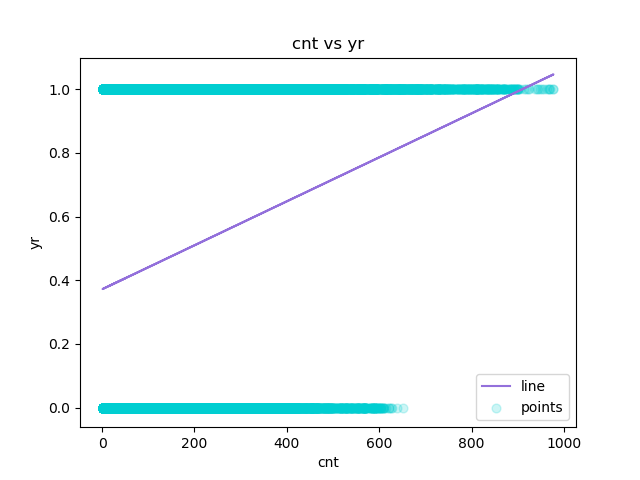


# Linear Regression

Running linear regression on this data set can show us some interesting increases and decreases in data between different plots. For example, the graph below tells us that there is a positive correlation between the amount of bike users using the temperature. The higher the temperature the more users.



Once again like in normal distribution since we can ploy any values as x and y columns we can come up with some illegible graphs. Like total users vs year. Since there are only two values for year (0=2011, 1=2012) we see the influx of users but the graph is very confusing at best. It takes a lot of effort to figure out what some graphs are trying to say since at first glance the line seems to be drawn randomly.



## Overall Conclusion

In conclusion, from the output I can see a few interesting correlations. Like the increase in users over time or how weather effects the use of the bike sharing services.

The ability to load in any data set is also very useful as allows for analysing most data available. This is great for a quick analysis but if you want to see some more detailed correlations algorithms like multiple polynomial regression would be better suited. But using such an algorithm requires customised code every time which takes extra time.

# Appendix 1: Reflective Learning Log

*Date:*

*Work Completed*

*Understanding Achieved:*

List:

* Commits on Oct 16, 2019
  + Added code to turn a csv file into a dictionary
* Commits on Oct 24, 2019
  + updated scaling for mobile to fix formatting issues
* Commits on Oct 26, 2019
  + added DataFrame class
  + Added support for loading in custom CSVs and printing them to cli
  + added menu for printing and processing data
  + Added TO DO statements, expanded menus to list all functions of program, added placeholder functions for the missing functionality.
* Commits on Oct 29, 2019
  + Added support for user specified column typing (int/float/string), added util fo check user inputs are as expected and added comments to functions inDataFrame class
  + Implemented detailed header printing
  + updated to dos
  + Added exception handling for string to integer conversions, Updated TextUtils with comments and more generic code
* Commits on Oct 30, 2019
  + Added 1st iteration of the linear regression algorithm with the abili ty to plot the graph
  + moved estimation functions to MathUtil and added support for export of graph, linear regression algorithm needs more testing
  + Fixed LR algorinth, added more files for export and changed directory to always optuput to ./output/date, tested dataset and found to interesting correlation (need to find a new one?)
* Commits on Nov 5, 2019
  + Updates based on user testing: More validation for inputs, changed gr aph colour scheme to make it more redable, fixed issue with graphs not clearing, fixed issue with datasets breaking on reload, also refactored duplicated code in main for maintainability
  + Updated comments in all files
  + Decreased code duplication and added ability to export all graph data at once, fixed some minor issues with selections too
  + Added graceful handling of FileIO for opening/writing files
  + Created constants file for different strings in the program, fixed issue with exporting all with default data types, organised directories and added custom delimeter support
  + Moved Utils into Utils directory
* Commits on Nov 8, 2019
  + Added most tests for ExceptionUtils
* Commits on Nov 11, 2019
  + updated directory structure and exceptions tests
  + Fixed paths to work with Python 3.7.4 (default, Aug 9 2019, 18:34:13) [MSC v.1915 64 bit (AMD64)] on win32
  + Added tests for the maths util, while testing figured out that slope and intercept formulas were the wrong way around, but still used correctly for calculating predictions
  + Added tests for text utils, fixed file names with underscores so pytest can now recognise them
  + file name change
  + Added tests for the data frame class, fixed issue where the invalid rows werent dropped correctly
  + Fixed some directory names
* Commits on Nov 14, 2019
  + Added some support for exporting normal distributions
  + Made adjustments to distributions

# Appendix 2: References

1. Fanaee-T, H. (2019). *UCI Machine Learning Repository: Bike Sharing Dataset Data Set*. [online] Archive.ics.uci.edu. Available at: https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset [Accessed 18 Nov. 2019].
2. GeeksforGeeks. (2019). *Linear Regression (Python Implementation) - GeeksforGeeks*. [online] Available at: https://www.geeksforgeeks.org/linear-regression-python-implementation/ [Accessed 18 Nov. 2019].
3. En.wikipedia.org. (2019). *Simple linear regression*. [online] Available at: https://en.wikipedia.org/wiki/Simple\_linear\_regression [Accessed 18 Nov. 2019].
4. Mathsisfun.com. (2019). *Normal Distribution*. [online] Available at: https://www.mathsisfun.com/data/standard-normal-distribution.html [Accessed 18 Nov. 2019].