

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Course** | **L9 Digital Transformation** | | | | | | | | |
| **Lecturer** | **Willo Roe** | | | | | | | | |
| **Subject:** | Analysis of Big Data [MSCDT-ABD/Dub/FT] | | | | | | | | |
| **Study Mode:** | Full time | **X** |  | | | Part-time | |  |  |
| **Assignment Title:** | Analysis of Big Data - Assignment 2025 | | | | | | | | |
| **Word count:** | 4-5 pages | | |  | **Weighting** | | 50% | | |
| Due date: | **Monday, 5th May 2025 via Moodle** | | | | | | | | |
| **Submit assignment to:** | **Assignments must be submitted through Moodle only (hard copy is not**  **required).**   * Submit an assignment [Cover Page](https://moodle.griffith.ie/mod/resource/view.php?id=206974) including your Name and Student No. This can be found on Moodle. * Sections 1-4 of your assignment must be submitted as a report in Microsoft Word format (no longer 5 pages in length) submitted on Moodle with both your first name and last name in the filename. | | | | | | | | |
| **Learning Outcomes addressed and assessed in this Assignment:**  (ii) critically analyse existing big data datasets and implementations, taking practicality, and usefulness of metrics into consideration;  (iii) investigate and demonstrate the role of statistics in the analysis of large datasets;  (iv) select and apply suitable statistical measures and analyses techniques for data of various structure and content, and present summary statistics; | | | | | | | | | |

# Assignment

# Analysis of Big Data [MSCDT-ABD/Dub/FT]

### Assignment Instructions

* You are analysing a dataset of Commuting times and modes within the canal cordon in Dublin. The data are generated by interviewing commuters on a single day about how long it has taken them to get to work.
* The provided [dataset](http://d1/) contains 281 rows, split across 4 modes of transport, namely Walking, Cycling, Bus and Car. **The full dataset is <here> and is also available to download from Moodle**
* The assignment is to interpret the dataset using Descriptive analytics, and build a Predictive model to predict the commuting time based on distance to travel.

### Dataset Columns:

* **Case Number**: **Unique** identifier for each commuter interviewed
* **Mode –** The method of travel
* **Distance**: Straight line distance from home to work
* **Time**: The actual time taken to travel to work in minutes

**Items to be addressed in your assignment:**

1. **Data Preparation**: Import and Clean the dataset. This may involve handling duplicates, removing incorrect entries (e.g. text values in numerical columns) etc**. Document all of the data cleansing steps applied to the dataset in your report**
2. **Exploratory Data Analysis (EDA)**: Before building the Linear Regression model, visually explore the cleansed data by building reports and a dashboard using **a spreadsheet or BI tool** to apply **Descriptive** analytics to the dataset.

Analyse your data **by Mode** as well the **overall dataset**, exploring **averages**, **spread, standard deviations and correlation coefficients** etc. in your submission.

Clearly highlight each of these findings in your final report, for example, what are the differences you observe when analysing the data by Mode, as opposed to analysing across the full dataset.

1. **Model Building & Evaluation**: Use JASP to build a **linear regression model** that predicts the **Travel Time** based on **Distance** for the full dataset.Evaluate the model's performance using appropriate metrics, such as Correlation coefficient (r), RMSE, Slope and Intercept etc;

What are your observations? What are the differences you observe when analysing the data by each Mode, as opposed to analysing across the full dataset?

Clearly explain the significance of the metrics that you have chosen to observe (R, R squared, RMSE etc), as illustrated in JASP, within the context of your model and expectation of accuracy and residual errors within your model.

* Use your linear regression models to determine an **equation for your regression line,** and use this equation to predict the following with 5-10 examples:
  + Predicted Travel Times based on distance across the full dataset with examples of inputs and outputs
  + Predicted Travel Times for the Cycle dataset only, as this may have a different correlation coefficient than other modes
  + Include the differences that you have observed when modelling the Cycle Regression model vs the full dataset Regression model.
  + Include visuals from JASP when documenting all of the above.

1. **Interpretation**: Create a report document (3-5 pages) of the data analysis and your findings. Include your written Interpretation of descriptive and predictive statistical findings for **Cycle data vs Full Dataset (Walk, Cycle, Bus and Car)** and summarise the statistical impact Distance has on Travel Times.

### Deliverables

**Report**: A written report (4-5 pages) that includes all steps of your analysis, from data preparation through to interpretation. Include screenshots of all **visualisations** from your analysis, and a discussion of your model's evaluation results.

Include screenshots from your Descriptive Analysis Predictive Analysis (JASP) in your report submission, containing text, visualisations and metrics including averages, spread and correlations.

**Additional instructions**

* Harvard referencing and citation must be used where required. [See here for more information](https://library.griffith.ie/writing-skills/referencing-styles/).
* You must submit a Griffith College cover sheet, found on Moodle (Student info page). You can download this cover sheet and then write your assignment below it.
* Only 1 document can be uploaded to each Moodle submission point.
* Breaches of academic integrity will not be tolerated. [See here for more information](https://library.griffith.ie/writing-skills/academic-integrity/).
* This is an individual assessment, students must not share their work with each other or submit work produced by another student.

**Academic Misconduct**

**As defined in QAE J6 Academic Misconduct Procedure**

**Plagiarism:** *Plagiarism includes copying or incorporating material derived from pre-existing work (published or unpublished) without the permission of the originator or without an established form of acknowledgement. It includes verbatim quotation, paraphrasing, imitation or other devices, which give the impression of being a learner’s original work. It also includes the exploitation of ideas from others without proper acknowledgement, which mostly occurs in research, project work or assignments.*

[See here for more information](https://library.griffith.ie/writing-skills/academic-integrity/avoiding-plagiarism/)

As per the procedure, the penalty for Minor Academic Misconduct, which has taken place for the first time: The learner is given zero in that element of the coursework that was proved to have resulted from academic misconduct. The learner must resubmit only that element of the coursework in which academic misconduct was detected. The result in that element of coursework shall be capped at 40%. If the learner does not resubmit the element within the prescribed timeframe, he / she shall receive a fail result for the entire piece of coursework. A record of the offence will be held internally.

* Examples of plagiarism in this assignment could include, but are not limited to, • Submitting a reference list that is not of your own making – copying an existing reference list from another source.
* Submitting a video that is not of your own making. This includes voice overs from other sources.
* Submitting work that has been completed by others – thesis factories, online help sites, collaboration with others.

### Marking Scheme

|  |  |  |
| --- | --- | --- |
| **Category** | **Description** | **Percentage** |
| **Data preparation and analysis** | Quality of data preparation, documented decisions made of data cleansing, data quality remediations. Work explained | 10% |
| **Understanding of descriptive analytics** | Demonstrated through visualisations and summarise data tables, graphs, use of appropriate tool to evaluate results. Clear explanation | 30% |
| **Understanding of linear regression concepts** | Demonstrated through the correct application of the models, and interpretation of results, Model accuracy, based on the chosen evaluation metric, use of JASP to evaluate results | 40% |
| **Report layout and writing quality** | Clarity of the report and organisation, readability, and documentation, clear use of visualisations to improve clarity of significant findings. Clear and straightforward language used. | 20% |

**Rubric**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Category** | **Excellent  (80 – 100%)** | **Very Good (70-79%)** | **Good  (60-69%)** | **Average  (50 – 59%)** | **Sufficient (40 – 49%)** | **Insufficient (30 – 39%)** | **Grossly Insufficient  (0 – 29%)** |
| **Data preparation and analysis 10%** | All duplicates, case number errors, missing values and typos cleaned. Working clearly explained | Almost all data issues addressed and explained | Majority of data issues addressed. Explanation substantially complete | A number of data issues missed, explanation addressed some issues | Most issues missed. Explanation attempted. | Some issues addressed. Explanation inadequate | All or almost all data issues missed |
| **Understanding of descriptive analytics 30%** | All appropriate EDA of entire dataset and subsets carried out. Bivariate EDA carried out. Use of tools and results explained clearly | Almost all EDA complete. Use of tools appropriate, with some gaps. Explanation helpful if not complete | Most EDA complete. Tools appropriate for the majority of analysis. Explanation almost complete. | EDA of full dataset, subsets and bivariate attempted. Partial understanding demonstrated | EDA of entire dataset OR of subsets OR bivariate completed. Some understanding demonstrated. | EDA attempted in part. Understanding attempted. | No or unclear EDA performed |
| **Understanding of linear regression concepts 40%** | Multiple models fitted using appropriate techniques to the entire dataset and to subsets. Clear observations on the models fit. Understanding of model terms and diagnostics demonstrated. Model applied appropriately to new data and explained | Regression almost complete. Good but incomplete observations and understanding demonstrated. Model application substantially complete. | Regression model complete for most cases requested. Some observations on model and some understanding demonstrated. Model application partially complete. | Regression model complete for at least one part of the dataset. Some understanding of model terms and diagnostics demonstrated. Model application attempted | Regression modelling attempted but not correct or convincing. Observations inappropriate. Model diagnostics not understood. Model application missing or inappropriate. | Incorrect model approach taken. No or incorrect observations. No clear diagnostics. Model application not attempted | No or inappropriate models proposed. Understanding of the linear regression process missing |
| **Report layout and writing quality 20%** | Report structure appropriate and aids understanding. Conclusions clear and appropriate. Exemplary use of graphics. Clear straightforward language used | Report structure clear but lacking some minor items. Almost all appropriate graphics included. Conclusions slightly unclear or incorrect. Language generally appropriate. | Report structure and content clear with one or two sections or headings missing. Most graphics supplied. Language clear for the most part | Report complete but lacking some sections or headings. Some use of graphics. Language sometimes confusing. | Report partially complete but confusing or unclear. Very limited graphics. Language unclear. | Report missing substantial content. No use of graphics. Language very unclear | Report missing, incomplete, unclear. |