# Milestone 1 Weight distribution and marking rubric

## Weight Distribution of the tasks:

EDA (20%)

| Task | % |
|---|---|
| Loading and basic exploration | 20 |
| Variety and complexity | 30 |
| Proper visualisation | 50 |

Cleaning(40%)

| Task | % |
|---|---|
| Detection | 20 |
| Handling | 50 |
| Checking and observing | 30 |

Transformation(30%)

| Task | % |
|---|---|
| Discretisation | 20 |
| Encoding | 30 |
| Scaling/Transformation | 35 |
| New features | 15 |

Code quality and overall notebook quality(10%) - Note that code quality will be checked on each task specifically and not just overall.

# Marking Rubric -

| Task/Mark | 0-20% | 20-50% | 50-75% | 75-90% | 90-100% |
|---|---|---|---|---|---|
| Loading and basic exploration | Csv file **not loading properly**, data is **not observed** or explored at all. | Csv file **loads properly** but **without proper indexing**, data is not observed or explored at all. | Csv loads properly and **indexed properly**. **Only 1** basic observation was made about the dataset. | Csv loads properly and indexed properly. **Few** methods used to explore and make multiple observations about the data. | Csv loads properly and indexed properly. **Multiple** methods used to explore and make multiple observations about the data. |
| Variety, complexity and accuracy of the insights found. | **No/1** question was asked to give further insight into the data. | Questions were asked but with **little to no** variety(i.e exploring just the distribution) and gave **incorrect insights**. | Type of Questions **varied more** and gave the **correct insight**. All questions were very **basic** and did not lead to deeper insight into the data.. | Questions varied and had **1 or 2** relatively complex questions that gave a deeper insight into the data. | Questions varied and had **multiple** relatively complex questions that gave a deeper insight into the data. |
| Proper and clear visual rep. and comments about the insights found. | **No** visual representation and no comments made about the insights. | **Misrepresen-tation** of the insights. **No** comments made. | **Misrepresen tation** of the insights (i.e the graphs used did not best represent your insights). Comments made on **few** of the insights. | Graphs were represented **properly** however they were **not properly labelled**. Comments made on **most** of the insights. | Graphs were represented clearly and were able to **effectively communicat e** the insights found. Comments made on **all** insights. |

| | | | | | |
|---|---|---|---|---|---|
| Detecting and observing unclean data | **No** detection of unclean data. | Unclean data were detected using **1** method only. **No** comments made about your findings. | Unclean data were detected/observed using **1** method only. Comments were made on **some** of your findings. | Unclean data were detected/Observed using **multiple** methods. Comments were made on **some** of your findings. | Unclean data were detected/Observed using **multiple** methods. Comments were made about **all** your findings. |
| Handling unclean data | Data was **not cleaned.** | Data was handled **improperly**. **No** justification of the technique used to handle the unclean data. | Data was handled **properly**. **Justification** made on the technique used to handle the unclean data. | Data Handled properly and justified. **Multiple** techniques were proposed for **some** of the unclean data. | Data Handled properly and justified. **Multiple** techniques were proposed for **all** types of unclean data handled. |
| Observing changes and checking data is cleaned | **Did not** check the data was cleaned. **No** observations made after handling the unclean data. | **Did** check the data was cleaned. **Few** observations made after handling the unclean data. No comments made on the observations/findings. | **Did** check the data was cleaned. **Few** observations made after handling the unclean data. **Few** comments made on the observations/findings. | **Did** check the data was cleaned. **Multiple** observations made after handling the unclean data. **Few** comments made on the observations/findings. | **Did** check the data was cleaned. **Multiple** observations made after handling the unclean data. Comments were made on **each** observation/findings. |
| Discretisat-ion | Dates **not** discretized. | Dates discretized **improperly**. | Dates were binned properly but with **incorrect labels.** | Dates were binned properly with **correct labels.** | Dates were binned properly with correct labels.**Comments/observations made** about the discretized dates. |

| | | | | | |
|---|---|---|---|---|---|
| Encoding | **No** encoding done. | Encoding was **performed** with **improper** techniques and no justification. | Encoding was performed with **proper** techniques and justification. | **Different** encoding techniques were introduced. Proper justification of the method chosen to encode. | Same as the 75-90 range + **Observations and comments** were made on the dataset after the encoding was performed(i.e how has it changed). |
| Scaling and/or Normalisation | No features were scaled or normalised. | **Incorrect** features were chosen to scale/norm. | **Correct** feature(s) were chosen to scale/norm. . **Proper** technique(s) used and **justified**. | Correct feature(s) were chosen to scale/norm. . **Different** techniques were introduced. Proper technique used for the feature(s) and justified. | Same as the 75-90 range + **Observations and comments** were made on the dataset after the encoding was performed(i.e how has it changed). |
| Additional features | **No** additional features created. | **1** additional feature created **improperly**. | **1** additional feature created **properly**. | **2** additional features created. **1** properly and **1** improperly. | **2** additional features created **properly**. |

| | | | | | |
|---|---|---|---|---|---|
| Load into new csv file | Data **not loaded** back into a csv file. | | | Data **loaded properly** but with **improper** naming. | Data **loaded properly** with **proper naming**. |
| Code and overall notebook quality | **Hard-coded** and **difficult to read/understand**. | Hard-coded. Code **could be understood** but with **improper variable naming**. | Code is generic and could be easily used for various datasets.Easy to understand.However, much of the **code was repeated.** | Code is generic and could be easily used for various datasets.Easy to understand. **Functions were created** for common tasks to **avoid repeating** writing the code for each task. | Same as 75-90 + Notebook is **structured** nicely and has a **clear** and **nice flow** to it. |