**PES University, Bangalore**

(Established under Karnataka Act No. 16 of 2013)  
 **UE19CS203 – STATISTICS FOR DATA SCIENCE**

**EVALUATION SCHEME**

**Assessment Policy: In Semester Assessment (ISA) Max: 60(will be scaled down)**

**Assessment Policy: End Semester Assessment (ESA) Max: 100 (will be scaled down)**

**Total =ISA + ESA = 100 Marks**

|  |  |  |
| --- | --- | --- |
| **Activity** | **Marks** | **Remarks** |
| ISA1 | 60 | Scaled Down to 25 |
| ISA2 | 40 | Scaled Down to 15 |
| Assignment | 40 | Scaled Down to 10 |
| **Total Marks** | **50** | |

|  |  |  |
| --- | --- | --- |
| **Activity** | **Marks** | **Remarks** |
| ESA | 100 | Scaled Down to 50 |
| **Total Marks** | **50** | |

**PES University, Bangalore**

(Established under Karnataka Act No. 16 of 2013)  
 **UE19CS203 – STATISTICS FOR DATA SCIENCE**

**PROJECT GUIDELINES**

* **Dataset Selection ( 5 Marks)**

You are encouraged to select your own dataset.

The requirements are,

* Dataset must have at least 500 or more observations and between 10 to 15 variables. **Done**
* The dataset’s variables should include categorical variables, discrete numerical variables, and continuous numerical variables. **Done**
* There should be at least 3 – 5% of missing values and NULL or NA.**Done**
* **Exploratory Data Analysis (7 Marks)**

Once you choose the dataset, do the following steps.

* Describe your dataset. Explain the meaning of the columns that is there in the dataset.
* Data Cleaning – Handle the missing data for both categorical and numerical variables (by dropping and imputing). **Done**

(Note: The missing values cannot be just ignored or deleted without examining)

* Remove unwanted observations – Duplicate/irrelevant/repetitive. **Done**
* Fix the typos and inconsistent capitalization. **Done**
* **Graph Visualization (7 Marks)**
* Visualize the dataset to exhibit meaningful insights from it. **Done**
* Use any three graph visualization techniques. **Done**
* Filter unwanted outliers. **Done**
* Numerical – Box plot / Histogram.
* Categorical – Bar chart.
* **Normalization and Standardization (5 Marks)**
* Compute the mean and variance for each of the columns. **Done**
* Normalize all the numeric columns, to make mean 0 and variance 1 **Done**
* Discuss why is normalization is needed? How does it affect dataset?
* Use graphs used to check whether the data is normal. **Done**
* **Hypothesis Testing (4 Marks)**
* State the research hypothesis.
* Perform statistical tests.
* Freedom to make your own hypothesis based on the columns.
* Decide whether the null hypothesis is supported or rejected.
* **Correlation (3 Marks)** **Done**
* Find the correlation between variables that are positively and negatively related.
* State inferences about it.
* **Presentation (5 Marks)**
* Presentation slides to be prepared.
* Screen shot of a particular part of dataset and related graphs can be put in PPT along with insights.
* **Report (5 Marks)**

A report of 4 to 5 pages needs to be submitted**.**

**Note: A team size of 4(Minimum number of students=Max. number of students=4) should be selected from your respective section. A Google sheet will be shared by your respective faculty to collect details about project.**

**Deadline for Project Details: 25th September 2020**

**Deadline for Project Completion: End of October 2020**

**Presentation Dates: 1st November 2020 to 10th November 2020.**