Data Exploration

1. Categorical
   1. Nominal – car type, gender
   2. Ordinal – ranking
2. Numerical
   1. Discrete – count pax
   2. Continuous – price, height, weight
3. Univariate
   1. Categorical
      1. Histogram
      2. Frequency Table
   2. Numerical
      1. Central tendency (mean, median, mode)
      2. Dispersion (min, max)
      3. Shape: Histogram, box, scatter
4. Feature Cleaning
   1. Missing values (why: model inaccurate, algorithm cannot work)
      1. How to handle missing

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Meaning | Pros | Cons |
| Deletion | All Missing values (rows) | Faster | Loose my records |
| Mean / Median/ Mode | Replace NA with mean/median/mode | Better than losing records | Distort Distribution |
| Random Value |  |  |  |

* 1. Outliers (why introduce Noise)
     1. How to handle

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Meaning | Pros | Cons |
| Deletion | All outliers |  | Loose outliers |
| Mean/median/mode | Replace outliers with mean/median/mode | Preserve Distribution | Loose outliers |
| Discretization | Convert continuous variable to discrete variable | Minimize the impact of outliers | Loose outliers |

1. Feature engineering
   * 1. Feature scaling (what: standardize the range of independent variable/normalization)
     2. Why >> algorithm to work fine.
2. How to handle?

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Meaning | Pros | Cons |
| Zscore | Remove mean scale all the data to unit variance  Z = x-x.mean/std | Range from 0 – 1  All columns scaled to 0 – 1 | Compress into narrow range |
| Min-Max scaling | Transforming the feature to one range, new column |  | Loose outliers |
| Robust scaling | Remove median scale all the data to unit variance  Z = x-x.median/ IQR |  |  |

1. Discretization (why: group some fields to Improve model performance

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Meaning | Pros | Cons |
| Equal width binning | (30-35, 35-45, 45-50) based on bins |  | Sensitive to skewness |
| Equal frequency binning | (30-35, 35-45, 45-50) based on bins frequency | Improvise | Loose data |

1. Feature encoding
   1. Convert the categorical fields/values to numerical

|  |  |
| --- | --- |
| Week | Topic |
| 1 | RAMPUP |
| 2 | BA |
| 3 | SQL |
| 4 | PBI |
| 5 | STATS |
| 7 | ML |