MAST30034: Applied Data Science

Assignment 1

Introduction

The aim of this project is to gain an initial insight into the New York City Taxi and Limousine Service Trip Record Data. The aim will be achieved through performing an initial analysis,

along with a visualization of the results.

In this project, by analyzing the datasets, we intent to understand the problem of” What will affect taxi drivers’ profit?”. We divide the problem into two aspects.

Aspect 1: What will increase the demand of taxi in New York?

Aspect 2: What will bring taxi drivers more tips in New York?

For aspect 1, we focus on the two majority factors that may increase the demand of taxi in New York, the sporting events and the weather.

For aspect 2, we consider that the date and time and the weather may influence the tips rates.

Data and Attribute Selection

The sporting event we chose is the 2018 US Open (tennis) which was

(4 marks)

Clearly states data period (1m)

Clearly states the three (or more) attributes to be analysed (1m)

Convincing justification for data period (1m)

Convincing justification for three (or more) attributes to be analysed (1m)

Pre-processing and Cleansing (4 marks)

Clearly states pre-processing and/or feature engineering steps (1m)

Clearly states data cleansing steps (1m)

Adequately investigates data for possible anomalies/outliers (1m)

Appropriate justification for pre-processing steps, as well as steps for handling missing data (1m)

Visualisation: Quality/Clarity (6 marks)

No marks possible without geospatial visualisation Geospatial visualisation is present (i.e heatmap, choropleth) (1m)

Appropriate granularity. Is it easily understandable to see what the visualisation is trying to show? Are there too many data points? (1m)

Geospatial visualisation clearly expresses a story, particularly if it raises “interesting” areas of further analysis or indicates an area that does not need further analysis (1m)

Appropriate choice of dimension, colour scheme, legend and formatting (1m)

Appropriate explanation of what the visualisation shows without being overly verbose (2m)

Analysis of result(s) (3 marks)

An appropriate summary statistic to describe the chosen attributes (1m)

Appropriate analysis of the relationship between the attributes (2m)

Quality and clarity of report (3 marks)

Quality and clarity of report (3 marks)

Quality writing, spell-checked, correct grammar, and comprehensible sentence structures (1m) Identifies potential stakeholders, motivation for the report and real-life use cases (1m) Provides recommendations for potential stakeholders based on analysis of findings (1m)