marshmallow

Data Science Task

We'd like to give you the opportunity to show off your skills by completing a short data exploration and modelling task, **in Python**.

Motor insurance prices vary according to many factors. One of the most important is the postcode where the insured vehicle is kept. This is also one of the most difficult to understand and model because there are so many (2 million+) separate postcodes.

For this task, you will investigate geographical variations in insurance pricing by examining data on a *postcode sector* level. That is, the postcode less the final two characters (e.g. EC1R 3). There are around 11,000 separate postcode sectors.

The aim is to use publicly available data on postcode sectors to gain insight into how the price of motor insurance relates to geographic location.

Task

1. Explore the provided data, process the data into features suitable for modelling and record any insights into patterns that you find.
2. Build a model to predict the price for postcode sectors. Split the data into appropriate training and test sets and build the best model you can to try and predict the *mean premium price* for a postcode sector. Evaluate the performance of your model.

You should return

- Your code
- A short report on your findings.

Please send back your work within **3 hours 10 minutes** of receipt of these instructions.

Hints

- Include comments/markdowns to explain your work as you go
- Don't worry about building the perfect model, or if you aren't able to complete everything – we are more interested in your approach to the task. Your code should allow us to see your work and understand your method!

Good luck!

Data specification

You have been provided with data in 3 separate CSV files. Each of the columns in each of those CSVs is explained here.

Dataset 1 - quote_prices.csv
One million motor insurance quote prices with the postcode. Prices are the cheapest price offered on a price comparison website.

Column	**Description**
premium_price	The total premium price for the quote
postcode	The postcode for the quote

Dataset 2 - postcode_sector_data.csv
Various publicly available data points for each postcode sector. These may correlate with the premium price in various ways.

Column	**Description**
postcode_sector	The postcode sector
relative_area	The relative area in km^2 of the sector
population_density	The mean population density of the sector
multiple_deprivation_index	The mean of the multiple deprivation index, a government index measuring deprivation
income_deprivation_index	A related government index for income
employment_deprivation_index	A related government index for employment
crime_deprivation_index	A related government index for crime
rural_urban	The category of area that constitutes the postcode sector
distance_to_station	The mean distance to nearest station for properties in the sector
never_worked	If the sector has a high or average proportion

	of residents who have never worked.
region	The wider region that the sector is in

Dataset 3 - local_authority_data.csv
Additional data that could relate to premium prices that is only available at the local authority level (less granular than postcode sector). It has been mapped to postcode sector appropriately.

Column	**Description**
postcode_sector	The postcode sector
road_usage	The road usage (in million vehicle miles) for the local authority that the postcode sector is in
total_offences	The total number of crimes in 2019 for the local authority that the postcode sector is in
vehicle_offences	The total number of vehicle-related crimes in 2019 for the local authority that the postcode sector is in

Further datasets
If you wish to use any additional publicly available data relating to postcodes that you think is relevant then feel free to do so.