HONG KONG INSTITUTE OF VOCATIONAL EDUCATION
DEPARTMENT OF INFORMATION TECHNOLOGY

Higher Diploma in Data Science and Analytics

ITP4887 Big Data Management

Assignment

Group Project

You are required to form a group of 2 classmates or even work individually. Each group or individual will receive one set of data. All groups or individuals have to process the dataset and generate some findings. The data is about Renting Post in the UK. All groups and individuals make use of provided data and finish the following tasks: Data pre-processing, Analysis, Reporting

In this case, a client employed you/your group to have a feasibility study on the circumstance of renting transactions in the UK, and your report is required to answer the following questions

1. Generate (5 to 10 rows) processed renting post records
2. What kind of property is having the most number of the bathroom on average?
3. What is the contribution of house type in the record?
4. What is the proportion of room facilities on average in the detached house?
5. Which date has the more renting post?
6. Is there any relationship between the number of bedrooms and the number of bathrooms?
7. Map question

Dataset Description

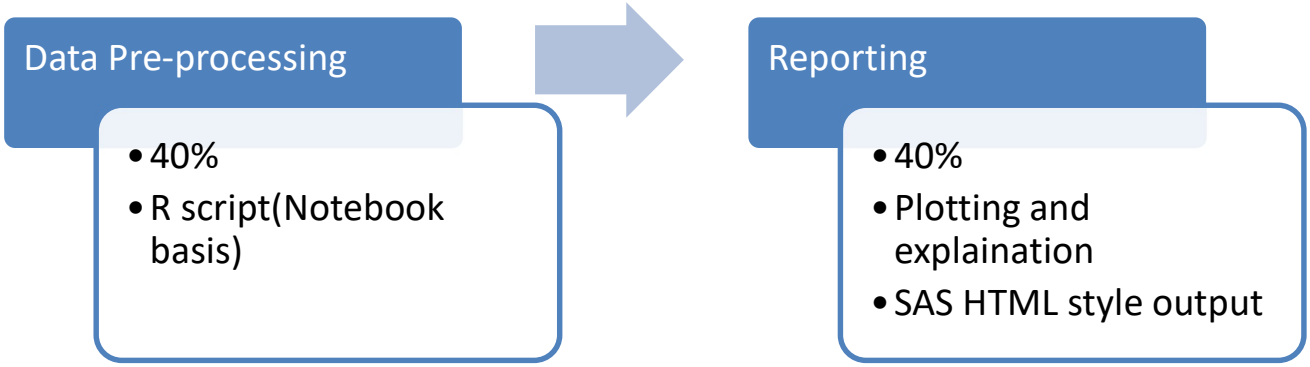
Renting Post from Zoopla.com

The dataset is from Zoopla.com, acquired by CMHO@Arical.ai. Indeed, Arial AI limited is an expert firm in geospatial analytics. The dataset has two days of renting post with the price, address, latitude, longitude, date, the number of facilities (includes bed, bath, reception[1]). The following is a sample of the dataset.

price	types	address	latitude	longitude	date	beds	baths	sq.	bath	reception	receptions	bed
865000	2 bed flat for sale	Pages Walk, London SE1	51.49574	-0.08120800	2nd Sep 2021	2	2	840	NA	NA	NA	NA
750000	3 bed detached house for sale	Park Road, Kenley CR8	51.32100	-0.10846600	2nd Sep 2021	3	NA	NA	1	1	NA	NA
419950	2 bed terraced house for sale	Anthony Road, London SE25	51.38741	-0.07190300	2nd Sep 2021	2	2	NA	NA	NA	2	NA
1095000	3 bed flat for sale	Pages Walk, London SE1	51.49574	-0.08120800	2nd Sep 2021	3	2	968	NA	NA	NA	NA
849995	4 bed semi-detached house for sale	Inchmery Road, London SE6	51.44025	-0.01340200	2nd Sep 2021	4	NA	NA	1	NA	3	NA
300000	2 bed flat for sale	North Block, The Railstore, Kidman Close, Gidea Park, R...	51.58480	0.21380000	2nd Sep 2021	2	NA	NA	1	1	NA	NA
625000	2 bed flat for sale	Holloway Road, London N7	51.55047	-0.10966300	2nd Sep 2021	2	2	NA	NA	1	NA	NA
1450000	3 bed flat for sale	Netherhall Gardens, Hampstead, London NW3	51.55218	-0.17717900	2nd Sep 2021	3	2	NA	NA	1	NA	NA
1750000	4 bed terraced house for sale	Offord Road, London N1	51.54364	-0.11171700	2nd Sep 2021	4	2	NA	NA	NA	2	NA

[1] a room in a house where people can sit together

Task 20% Project Management by Github +80% Tasks



Project Management

Your group should nominate a representative to create a private Github server and name the repository as "RepresentativeName_4887Project", your group have to invite account bensonlau@vtc.edu.hk as a collaborator. The Repository should include all deliverables such as R notebook and SAS script. **Higher marks will award to groups that have a regular and reasonable update to their project with change descriptions.**

Criteria

1. Private Github Server (5M)
2. Invite the target as a collaborator (5M)
3. Document in Repository (5M)
4. change description (5M)

Data Pre-processing

Your group are received 10000 Renting Post for your project, and you have to pre-process data using R before you pour the data to SAS:

1. Convert the data source format (JSON) to the dataframe. (2M)
2. Tidy up the date format to become readable, i.e., 2nd Sep 2021 => 2/9/2021. (15M)
3. Fill all missing data with 0s except the sq. column. (5M)
4. Aggregate the number of corresponding facilities (Bed, Baths, Reception). (5M)
5. Extract Property type from "types" column. (10M)
6. Output your pre-processed data to CSV format called "Project_Housing.csv". (3M)

price	types	address	latitude	longitude	date	sq.	TotalBeds	TotalBaths	TotalReceptions	FlatType
865000	2 bed flat for sale	Pages Walk, London SE1	51.49574	-0.08120800	2020-09-02	840	2	2	0	flat
750000	3 bed detached house for sale	Park Road, Kenley CR8	51.32100	-0.10846600	2020-09-02	NA	3	1	1	detached house
419950	2 bed terraced house for sale	Anthony Road, London SE25	51.38741	-0.07190300	2020-09-02	NA	2	2	2	terraced house
1095000	3 bed flat for sale	Pages Walk, London SE1	51.49574	-0.08120800	2020-09-02	968	3	2	0	flat
849995	4 bed semi-detached house for sale	Inchmery Road, London SE6	51.44025	-0.01340200	2020-09-02	NA	4	1	3	semi-detached house
300000	2 bed flat for sale	North Block, The Railstore, Kidman Close, Gidea Park, R...	51.58480	0.21380000	2020-09-02	NA	2	1	1	flat
625000	2 bed flat for sale	Holloway Road, London N7	51.55047	-0.10966300	2020-09-02	NA	2	2	1	flat
1450000	3 bed flat for sale	Netherhall Gardens, Hampstead, London NW3	51.55218	-0.17717900	2020-09-02	NA	3	2	1	flat
1750000	4 bed terraced house for sale	Offord Road, London N1	51.54364	-0.11171700	2020-09-02	NA	4	2	2	terraced house
465000	2 bed flat for sale	Purbrook Estate, Tower Bridge Road, London SE1	51.49867	-0.07875100	2020-09-02	NA	2	1	1	flat

Deliverable:

1. One Dataset File in .csv format
2. R notebook

Reporting

Base on the client's concern, you may visualise the data for answering them. You are required to upload your pre-processed dataset to your SAS OnDemand Server. You may answer the questions with a graphic and some descriptive comments. You have to consolidate all answers to a single report in SAS HTML Output format. There is a summary of the client's concern:

Five Row of Data form dataset

Obs	price	types	address	latitude	longitude	date	sq.	TotalBeds	TotalBaths	TotalReceptions	FlatType
1	865000	2 bed flat for sale	Pages Walk, London SE1	51.495738	-0.081208	2020-09-02	840	2	2	0	flat
2	750000	3 bed detached house for sale	Park Road, Kenley CR8	51.320999	-0.108466	2020-09-02	NA	3	1	1	detached house
3	419950	2 bed terraced house for sale	Anthony Road, London SE25	51.387405	-0.071903	2020-09-02	NA	2	2	2	terraced house
4	1095000	3 bed flat for sale	Pages Walk, London SE1	51.495738	-0.081208	2020-09-02	968	3	2	0	flat
5	849995	4 bed semi-detached house for sale	Inchmery Road, London SE6	51.440248	-0.013402	2020-09-02	NA	4	1	3	semi-detached house

The datasetprinted by print proc in SAS

1. Generate (5 to 10 rows) processed renting post records. (5M)
2. What kind of property is having the most number of the bathroom on average? (5M)
3. What is the contribution of house type in the record? What is the most common type of property in the UK? (5M)
4. What is the value distribution of the number of Reception between the semi-detached house and terraced house? (10M)
5. What kind of property is contain the second most turnover? (5M)
6. Is there any relationship between the number of bedrooms, the number of bathrooms and the average price of a different property? (10M)

Submission Day

Section	Date	Deliverable
Final Submission	Check Moodle	R script SAS Script SAS Report HTML format