Airbnb Lab

Please read all of the guidelines carefully before submitting the lab.
☺ There are ***200 points*** in total.

Due date: Friday, March 5, 11:59 PM. Late submissions will be accepted with a penalty!

Deliverables:

1) The code of the project in **.py** or **.ipynb** format (in several files, as explained below)
2) The lab report written with ***LaTeX*** and exported in **.pdf** format (one file)

Guidelines – Before You Start

1) You will be using the **Python** programming language for this lab. You need to write your codes in an empty **.py** OR **.ipynb** file. Don't use the files we used for labs to submit your homework or other assignments (those submissions will not be graded.).
2) Make sure that you provide many comments to describe your code and the variables that you created.
3) Please use the ***IEEE*** journal template on ***overleaf.com***. Here is the link: https://www.overleaf.com/latex/templates/preparation-of-papers-for-ieee-sponsored-conferences-and-symposia/zfnqfzzzxghk
   To be able to work on ***overleaf.com***, you will need to register first (you can also compile your ***LaTeX*** file locally.)
4) For some of the code, you may need to do a little bit of "**Googling**" or review the documentation.

What is Airbnb?

Airbnb, Inc. is an online marketplace for arranging or offering lodging, primarily homestays, or tourism experiences. The company does not own any of the real estate listings, nor does it host events; it acts as a broker, receiving commissions from each booking. The company is based in San Francisco, California, United States.

The company was conceived after its founders put an air mattress in their living room, effectively turning their apartment into a bed and breakfast, in order to offset the high cost of rent in San Francisco; Airbnb is a shortened version of its original name, AirBedandBreakfast.com.

Interesting Airbnb houses from different cities (left to right): Van Gogh's room in Chicago, Hobbit house in Thailand, Super Mario Bros house in Tokyo, an igloo in Brooklyn, New York.

Lab - Coding

In this assignment, you will be working with the Airbnb data collected from Boston, Massachusetts. This dataset provides information on the Airbnb listings in Boston. The data has been provided in three different datasets: (i) **calendar.csv**, (ii) **listings.csv**, (iii) **reviews.csv**. **listing_id** is a column that is common to all three datasets, and it is a unique number assigned to a place to stay advertised on Airbnb.

Open the Excel files and take a look at them before starting. The names of the columns should be self-explanatory. For any questions, please refer to the link where the data was downloaded from (Kaggle: https://www.kaggle.com/airbnb/boston) or send an e-mail.

Before you start with the questions below, create a new and empty folder called **data_350_airbnb_lab**. Call the file where you will write your code **data_350_airbnb_lab**, as well.

Part I: Setting Up Your Dataset and Descriptive Statistics (10 points)

Q1: Import the three datasets by using the ***read_csv( )*** function of *Pandas*. Name the datasets that you imported as **calendar_df**, **listings_df**, and **reviews_df**. Add brief comments to your code and explain what your code is doing (Do not drop the NA values) (**5 points**).

Q2: Take a look at the numeric variables in your **listings_df** and datasets. Provide a descriptive table that shows the ***minimum***, ***maximum***, ***mean***, ***median***, ***variance***, and ***standard deviation*** values for each variable. Here is a template:

Variables	Minimum	Maximum	Mean	Median	Variance	Std. Deviation
Var1						
Var2						
…						

A list of the variables that you need to look at:
host_response_rate, host_acceptance_rate, host_listings_count, host_total_listings_count, accommodates, bathrooms, bedrooms, beds, price, weekly_price, monthly_price, security_deposit, cleaning_fee, guests_included, extra_people, minimum_nights, maximum_nights, availability_30, availability_90, availability_365, number_of_reviews, review_scores_rating, review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_value, reviews_per_month.

Is there anything "strange" about the dataset? Are there any values that seem to be badly measured? (The answer to this question does not need to be "**yes**". You just need to elaborate.) (**5 points**)

Make sure that all of your code is running!

Open an empty file in a text editor (like *Notepad++* on Windows or *Atom* on Mac). Copy and paste the code (along with your comments) into the empty file. Go to the "Save As…" option and save the code file you have created as "**part_1.py**" in the **data_350_airbnb_lab** folder you have created at the beginning.

Part II: Sentiment Analysis and Adding New Data (30 points)

In this question, you will be doing two types of sentiment analysis to convert the reviews into a numerical score of "**positivity**" and "**negativity**". Sentiment analysis is a simple Natural Language Processing (NLP) technique that allows you to quantify the sentiment in a text dichotomously.

The sentiment analysis will assign the reviews with more "positive" comments a greater ***positivity*** score, and the reviews with more "negative" comments will receive a greater ***negativity*** score. A template for doing the sentiment analysis can be found in the assignment folder. The file name is **sentiment_analysis_template.py**.

For this assignment, you will need to install the **nltk** package of Python by using the "***pip install nltk***" or "***conda install nltk***" commands in your terminal / command prompt.

Q3: Take a look at the **comments** column of the **reviews_df** dataset. Run the sentiment analysis on each cell in the **comments** column and add four new columns to your dataset as a result: **negativity**, **neutrality**, **positivity**, **compound** (**10 points**).

Q4: Now, you will work with your own sentiment analysis technique! Take a look at the list of positive (**positive_words.csv**) and negative words (**negative_words.csv**) in the assignment folder. Write a loop structure that iterates over each comment in the **comments** column of the **reviews_df** dataset. Count the following in each comment: (**i**) total number of words, (**ii**) total number of positive words, (**iii**) total number of negative words. And create the following variables:

positivity_simple = total number of positive words / total number of words
negativity_simple = total number of negative words / total number of words

Based on what you calculated, add two new columns to your **reviews_df** dataset: **positivity_simple** and **negativity_simple** (**10 points**).

Q5: Now, you will add the **means** of the scores that you calculated in **Q4** to the **listings_df**. Finding the unique values of listings in the **reviews_df** dataset (**listings_id** column). Calculate the average scores for each listing and name them in the following way: **negativity_mean**, **neutrality_mean**, **positivity_mean**, **compound_mean**, **positivity_simple_mean**, **negativity_simple_mean**. Add these values to the **listings_df** dataset as new columns (**10 points**).

Make sure that all of your code is running!

Open an empty file in a text editor (like *Notepad++* on Windows or *Atom* on Mac). Copy and paste the code (along with your comments) into the empty file. Go to the "Save As…" option and save the code file you have created as "**part_2.py**" in the **data_350_airbnb_lab** folder you have created at the beginning.

Part III: Linear Regression (60 points)

Q6: We need to develop a model that predicts the **price** (**Y**) column based on the numeric variables we had and the sentiment scores we calculated. For this part, we will be focusing on the following explanatory variables (**X**'s):

host_response_rate
review_scores_rating
review_scores_accuracy
review_scores_cleanliness
review_scores_checkin
review_scores_communication
positivity_mean
negativity_mean
positivity_simple_mean
negativity_simple_mean

Develop a linear regression model that uses price as the outcome variable (**Y**) and all the other variables listed above as explanatory variables (**X's**).

To run the linear regression, please take a look at the **linear_regression_lab.py** file in the assignment folder. Using different variations of cross-validation and using a 70-30 split ratio, run the linear regression model and report the coefficients for your variables. You will be using **train/test split CV**, **K-fold CV (where k=5)** and **LOOCV** (**20 points**).

Q7: For this question, manually code the following cost functions: **R^2**, **MSE**, **Root MSE**, **MAPE**. Using the results from **Q6**, create a table that shows different error rates corresponding to different types of cross-validations and cost functions. Below is an example (**40 points**).

	R^2	*MSE*	*Root-MSE*	*MAPE*
Train-Test CV				
K-Fold CV				
LOOCV				

Make sure that all of your code is running!

Open an empty file in a text editor (like *Notepad++* on Windows or *Atom* on Mac). Copy and paste the code (along with your comments) into the empty file. Go to the "Save As…" option and save the code file you have created as “**part_3.py**” in the **data_350_airbnb_lab** folder you have created at the beginning.

Part IV: Principal Components Analysis (30 points)

Q8: From the regression coefficients (X’s) we used in **Part III**, it looks like there are three main groups of coefficients that determine the price (at least in theory): **host response rate**, **review scores**, and the results of the **sentiment analysis**. Now, we will do PCA to see if we (should) indeed get three categories after the dimensionality reduction process.

Here is one resource for how to do **PCA** on **Python**: https://www.geeksforgeeks.org/principal-component-analysis-with-python/

For this part of the analysis, please do the following:

- Split the dataset you used for linear regression into all X’s and Y (explanatory variables and outcome variable)
- Split the **X** and **Y** datasets in training and test sets
- Standardize the values by using the **StandardScaler()** function of **sklearn** package
- Apply the PCA function of **sklearn** into training and testing sets, set **n_components = 3**
- Fit linear regression to the training set and **report your coefficients**
- Predict the test set result
- Calculate the **R^2**, **MSE**, **Root MSE**, and **MAPE** values
- Compare the error values to the first row of the table you created in **Q7**
- Compare the performance of **PCA** to linear regression. Which one do you think did a better job?

Report the coefficient values, error rates, and your comments (**30 points**).

Make sure that all of your code is running!

Open an empty file in a text editor (like *Notepad++* on Windows or *Atom* on Mac). Copy and paste the code (along with your comments) into the empty file. Go to the “Save As…” option and save the code file you have created as “**part_4.py**” in the **data_350_airbnb_lab** folder you have created at the beginning.

Part V: Visualization (20 points)

Q9: For this part of the analysis, you can use any visualization tool within the Python realm. Some useful packages are *matplotlib* and *seaborn*. You can find some helpful documentation here:

https://matplotlib.org/
https://seaborn.pydata.org/

i. Take a look at the compound column you created in the sentiment analysis. Report the percentage of listings that have a **compound_mean** below zero and those that have a **compound_mean** above zero using a bar chart or a pie chart (**5 points**).
ii. Create a "correlogram" that shows the correlations between the numerical variables you have in **Q6.** Include the trend lines, as well. (*Hint*: an easy option would be the **pairplot** provided in the **Seaborn** package) (**5 points**).
iii. Create a pairplot that shows the relationship between the three principal components you created in Part IV (**5 points**).
iv. Create a table for your linear regression output that shows the coefficients for different variables, number of observations, degrees of freedom, statistical significance, R^2 value etc. (**5 points**).
An example can be found on Page 3 here: https://cran.r-project.org/web/packages/stargazer/vignettes/stargazer.pdf
v. Add the tables you created in **Q2** and **Q7** to your set of visuals.

Make sure that all of your code is running!

Open an empty file in a text editor (like *Notepad++* on Windows or *Atom* on Mac). Copy and paste the code (along with your comments) into the empty file. Go to the "Save As…" option and save the code file you have created as "**part_5.py**" in the **data_350_airbnb_lab** folder you have created at the beginning.

Part VI: Creating the lab report (50 points)

Q10: Write a report (2 pages) that includes all of your findings and the visuals that you created. The report that you write should use the *IEEE format* and include the following sections:
Abstract: A short summary of your report
Introduction: A summary of what you expected and did, and two-three of your most significant findings
Data: A description of your dataset (number of observations, what kind of information is there, descriptive statistics table etc.)
Results: A summary of the analyses you completed for the lab + a discussion on the ***results*** that talks about these topics:

- What are some of the main findings you have?
- Which variable(s) seem(s) to explain the price the most?
- What would be some of the ways to improve this analysis?
- Do you see any typical causal inference problems (***selection bias***, ***simultaneity***, ***omitted variable bias***) in this analysis?

Final step:
Send what you saved in your submission folder **(data_350_airbnb_lab)** through *Notebowl*.