IE 360 Statistical Forecasting and Time Series
Homework 1, due March 29th, 2024

Data

Central Bank of the Republic of Turkey (CBRT) provides several economy related measures on their Electronic Data Delivery System: https://evds2.tcmb.gov.tr/ . If you click "All Series" link on the upper menu, you will see the data categories in an alphabetical order on the left. Figure 1 shows some part of this menu. Once you click one of the categories, you will be able to see the available data under the corresponding category. Figure 2 provides an example for "Housing and Construction Statistics".

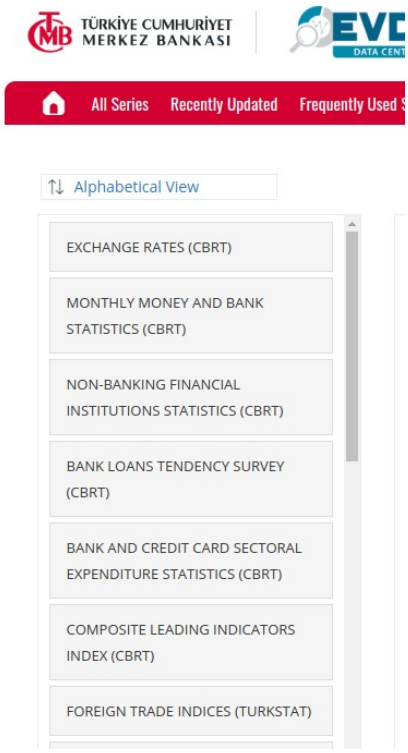

Figure 1. Series categories from Electronic Data Delivery System on https://evds2.tcmb.gov.tr/

Data Requirements

The aim of this homework is to provide a demonstration of your time series data manipulation and regression skills on a selection of time series data from the provided link. You are free to choose the type of the series you are willing to work on as long as your selection has the following properties:

1. There should be series from at least 3 different categories of data (i.e. Exchange Rates, Price Indices and Employment Statistics)
2. Times series from each category mentioned in Requirements 1 should not be correlated. Try to find out three time series where the absolute correlation coefficient between each pair is less than 0.5.
3. The observations are sampled at most monthly level (i.e. frequency of the observations is at most monthly)
4. The length of the period for the observations is at least 2 years for each series.

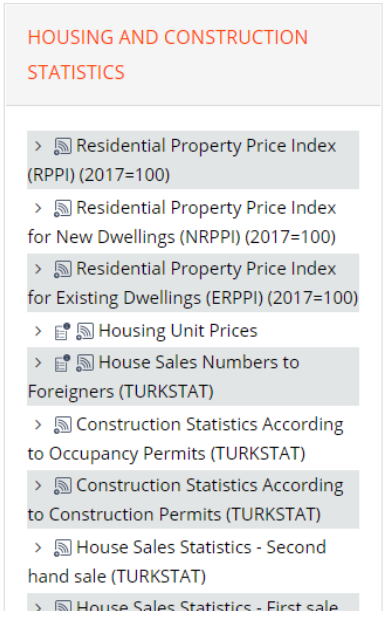

Figure 2. Series available under “Housing and Construction Statistics” category

Google Trends as an alternative data source

Another interesting data source is Google Trends (http://trends.google.com/). The Wikipedia definition for Google Trends is:

“Google Trends is a public web facility of Google Inc., based on Google Search that shows how often a particular search-term is entered relative to the total search-volume across various regions of the world, and in various languages. The horizontal axis of the main graph represents time (starting from 2004), and the vertical is how often a term is searched for relative to the total number of searches, globally.[1] Below the main graph, popularity is broken down by countries, regions, cities and language. Note that what Google calls "language", however, does not display the relative results of searches in different languages for the same term(s). It only displays the relative combined search volumes from all countries that share a particular language (see "flowers" vs "fleurs"). It is possible to refine the main graph by region and time period.”

Sometimes it is interesting to check if search volume of a certain keyword has some relation to an event. For example, number of searches for “flu” can give you an idea about the number of people affected by the flu over the world. Another example is from a senior design project we have worked on last year. It was about understanding if Google Trends provide information about the number of pay-per-view (PPV) customers of Digiturk. To be more specific, Lig TV is a football channel of Digiturk holding rights to broadcast Turkish Super League live since 2001. It serves its customer through different channels such as the Internet platform “Digiturkplay” or satellite TV “Digiturk Plus”. All channels include a service called pay-per-view (PPV) providing customer to pay for the particular programs that they watch via private telecast. The aim here is to see if it is possible to forecast demand for PPV through the Internet search data, namely Google Trends Data. To achieve this, we obtained PPV data from Digiturk. For each soccer game, we obtained the daily search volume information for each team playing a game (i.e. if the game is between Beşiktaş and Fenerbahçe, we obtained the search volume for each team on the game day and sum these values to find out a measure that will quantify the popularity). The plot of the PPV sales versus the search volume information for each game is schematized in Figure 3. As you can see, Google Trends data is highly correlated with the sales.

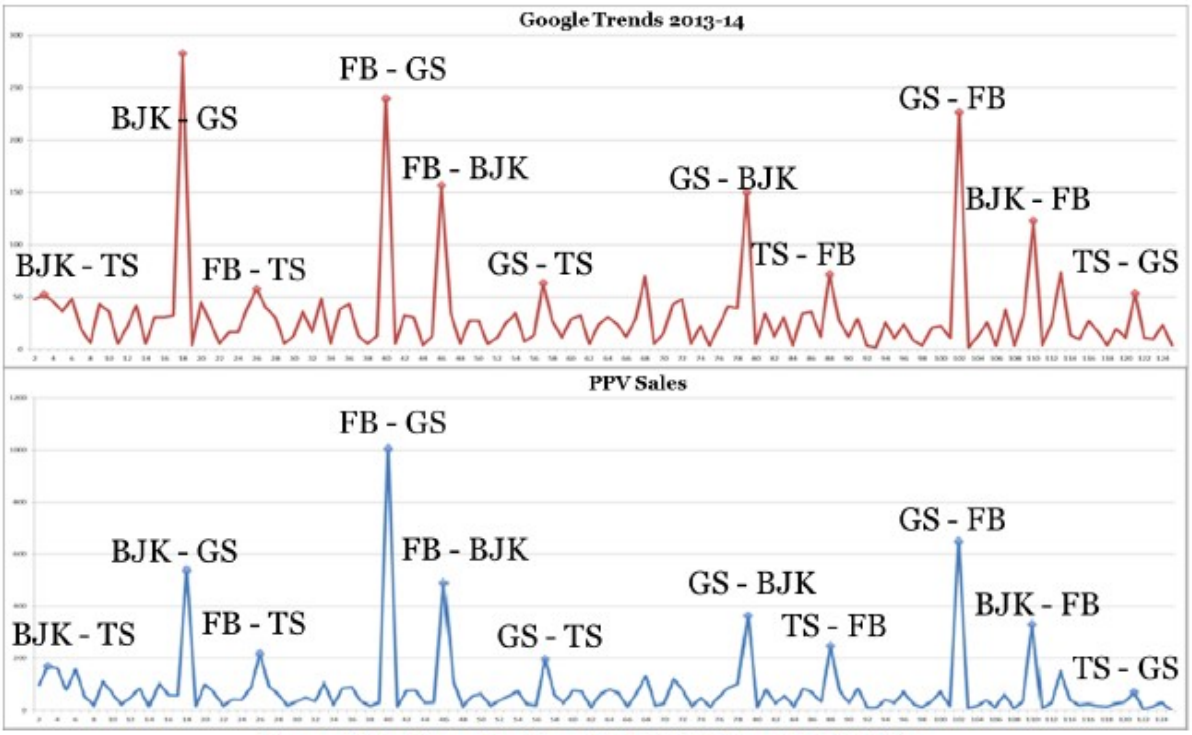

Figure 3. Search volume versus PPV sales

Analyses and Evaluation

You are expected to make reasonable series selections from CBRT and relevant Google Trends data so that your visualizations and analyses make sense. To be more clear, the selected data should not be irrelevant. You can start with a relevant research question such as: "Are house sales statistics related to housing unit prices, exchange rates, consumer price index and etc.?" Any relevant analysis will be accepted as long as you clearly motivate your purpose. Please note that this is just to show your time series regression skills together with data manipulation and visualization.

<u>For the time series data manipulation and visualization, the following analyses are required:</u>

Determine a search keyword that can be relevant to the measures you consider (i.e. if you are interested in unemployment rate, you can check the search volume for "iş ilanı"). Plot the search volume and the data you obtained (consider potentially relevant visualizations, your choice of the visuals is part of the homework assignment). Note that you need to provide these plots for each measure of interest. Is there any trend/relation you can observe?

<u>For the time series regression, the following analyses are required:</u>

You will be building time series regression models for each category mentioned in Requirements 1. You should try various regression models and come up with a model that explains sales in terms of independent (i.e. Google trends data or others) and trend/seasonality variables. You should present all the statistical evidence supporting the validity of your analysis (residual analysis, significance of coefficients, etc.). This should be done only for your selected model.

You will be evaluated based on the following criteria:

- Format (10 points)
- Soundness of your research question (20 points) – An introduction section
- Data manipulation and visualization (20 points) – An analysis section where you read and merge the data, create time series objects, visualize data, find correlations and comment based on your research question and findings.
- Time series regression analyses (40 points) – An analysis section where you build time series regression models with each CBRT category as target (i.e. Three time series regression models will be built).
- Concluding remarks (10 points) – A conclusion section

Report & Code Documentation:
Consolidate your results, visual aids included, into a detailed report. Wrap up your report with the outlined criteria in the previous section.

Please solve the following exercises using R (http://www.r-project.org/) or Python (https://www.python.org/). You are expected to use GitHub Classroom and present your work as an html file (i.e. web page) on your progress journals. There are alternative ways to generate an html page for you work:

a) A Jupyter Notebook including your codes and comments. This works for R and Python, to enable using R scripts in notebooks, please check:
    a. https://docs.anaconda.com/anaconda/navigator/tutorials/r-lang/
    b. https://medium.com/@kyleake/how-to-install-r-in-jupyter-with-irkernel-in-3-steps917519326e41

    Things are little easier if you install Anaconda (https://www.anaconda.com/). Please export your work to an html file. Please provide your *. ipynb file in your repository and a link to this file in your html report will help us a lot.

b) A Markdown html document. This can be created using RMarkdown for R and Python. Markdown for Python

Note that html pages are just to describe how you approach to the exercises in the homework. They should include your codes. You are also required to provide your R/Python codes separately in the repository so that anybody can run it with minimal change in the code. This can be presented as the script file itself or your notebook file (the one with *.ipynb file extension).

The last and the most important thing to mention is that academic integrity is expected! Do not share your code (except the one in your progress journals). You are always free to discuss about tasks but your work must be implemented by yourself.

Please mention if you utilize Large Language Models (i.e. chatGPT, BART and etc). You are expected to provide your prompt and resulting response of these models at the end of your reports as Appendix if you have utilized any of these tools.

As a fundamental principle for any educational institution, academic integrity is highly valued and seriously regarded at Boğaziçi University.

Best of luck, and remember – the journey of analyzing is as essential as the conclusions drawn!