INFS5710 Information Technology Infrastructure for Business Analytics

Project Statement

(Due by noon 12 PM on Monday 21 November 2022 via Moodle)

- **This project accounts for 30% of the total marks for this course.**
- **The deliverable is a PowerPoint file with video narration and speaker notes, and an appendix file.**

Bike sharing has become increasingly popular across the globe. Today, such programs operate in more than 1,000 cities, with more than half a million bicycles in use. The principle of bike sharing is simple: individuals use bicycles on an as-needed basis without the costs and responsibilities of bike ownership. It is short-term bicycle access, which provides its users with an environmentally friendly form of public transportation. This flexible scheme targets daily mobility and allows users to access public bicycles at unattended bike stations; bicycle reservations, pickup, and drop-off are all self-service. Commonly concentrated in urban settings, bike sharing programs also provide multiple bike station locations that enable users to pick up and return bicycles to different stations.

This project is about Capital Bikeshare (CaBi) in the metropolitan area of Washington DC (DC), which covers not only the DC area, but also some parts of two nearby states, Maryland (MD), and Virginia (VA). You are a business consultant working for the bike-sharing program.

Bike-sharing data

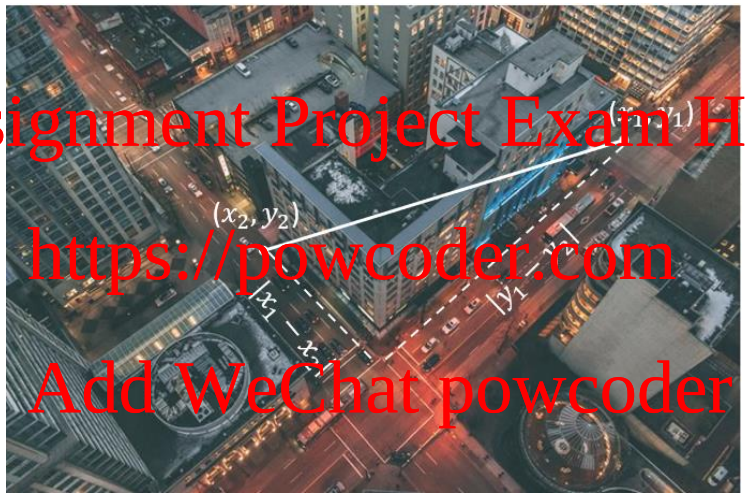
Your manager just referred you to download historical bike-sharing data by first visiting the following site https://ride.capitalbikeshare.com/system-data, then click "downloadable files". This would direct you to the following site https://s3.amazonaws.com/capitalbikeshare-data/index.html, which contains data of millions bike trips from July 2010 – 2022 September. Since the data come from the US, please be aware of the difference in date formats between the US (mm/dd/yyyy) and Australia (dd/mm/yyyy). It is also known that CaBi has changed the format of the data files recently. It is part of this project that you need to decide how to consolidate tables coming from different sources and/or with different formats.

Regional factors

As said, CaBi not only serves DC, but some cities in MD and VA. Even within DC, the district is divided into four quadrants of unequal areas: Northwest (NW), Northeast (NE), Southeast (SE), and Southwest (SW). Each city and DC quadrant presents distinct characteristics (e.g., some are culturally rich, some are more populated, and some have more crimes). Therefore, different regions may reveal different bike-sharing use patterns. You may download detailed information of all CaBi bike stations from https://opendata.dc.gov/datasets/capital-bike-share-locations/, in which the last column (attribute) REGION_NAME shows whether a station is in DC, VA, or MD. If a station is within DC, the attribute NAME would reveal the corresponding quadrant that it is located.

In the above file for station locations, you can find the locations of bike stations in the GPS coordinate system. For example, the coordinate of a station is (x, y), where x is the longitude coordinate and y is the latitude coordinate. The following link helps you to understand more about the GPS coordinate system: https://www.ubergizmo.com/how-to/read-gps-coordinates/. If you want to locate a place on Google Map by its latitude and longitude, you can also do it. For details, see the following link https://support.google.com/maps/answer/18539.

If you are interested in estimating the distance traveled for a ride, assuming that a bike rental starts from (x_1, y_1) and ends at (x_2, y_2), it is recommended that you estimate it using the so-called taxicab distance, which is $|x_1 - x_2| + |y_1 - y_2|$. See the following figure for interpretation. For more information, please see https://study.com/academy/lesson/taxicab-geometry-history-formula.html. Note that whether the distance is in degrees (without any conversion from longitude-latitude coordinates), miles, or kms, your findings, interpretations or insights should not change. It is recommended that you just quote the distances in degrees.



Weather data

Weather plays an important role when people decide whether or not to use bike-sharing. You are required to explore the relationship between weather (e.g., temperature, wind speed and humidity) and the bike-sharing rentals in this project. There are two known ways to download free historical weather data. The first way is to manually capture weather data month by month from Weather Underground (wunderground.com).

- First visit https://www.wunderground.com/ and try to search the weather condition in DC. (There are other locations in MD and VA that you may also try, where there are many bike stations as well.)
- You will be led to the site of a nearby weather station, which may be different from time to time.
- Click the History tab on the page, and then choose to view Monthly weather data. Once you choose a month, click View. For example, the following link shows the weather data of Oct 2011

measured at the Ronald Reagan Washington National Airport station (within DC): https://www.wunderground.com/history/monthly/us/va/arlington/KDCA/date/2011-10

- Scroll down the page, and you will see the table of Daily Observations. Use your mouse to copy the table and paste it to an Excel spreadsheet.
- Copy only the data required, i.e., July 2011 – 2022, for this project.

Another way, as suggested by a former student, is to download weather data from NOAA. Try: https://www.ncdc.noaa.gov/cdo-web/search. Search for "Daily Summaries" at relevant weather stations for a time period then "Add to Cart" - NOTE that this is a free service, but you'll have to type in an email address so that you can get the data download link once it processes.

Holiday data

Another factor that influences the bike-sharing rentals is holidays. You can easily search the dates of the US federal holidays and/or MD, and VA state holidays each year.

The Task

Your manager asked you to collect and analyze the data and "let the data speak". You understand that the company wants to further grow the market and induce more users. Before they do it, they want to have some insights from the data.

In this project, you are expected to manage and clean the data collected; some of them may contain missing data, different formatting, and incomplete information. The goal is to overcome such obstacles commonly encountered in reality to derive business insights from the datasets that can be used to promote CaBi's bike-sharing business.

Borrowing the terms from Data Warehouse, the following are some "dimensions" for the analysis in this project: station(s), time (including holidays), weather, membership, region, and bike-type. There is one obvious "measure" in this context, which is the bike use, the number of rides, or the demand. We define an "analysis topic" as one that studies how a measure changes according to one or multiple dimensions. For example, you may study the daily demand pattern and how it changes over the past 10 years, under different weather conditions and/or whether the day is a holiday. In this example analysis topic, weather data and holiday data are utilized. Note that you need to make sure that an analysis topic must be meaningful.

For this project, you are expected to choose no more than three analysis topics to study. It is more preferred that you study one topic in depth, rather than multiple ones superficially. **There are two constraints for your study**: (i) You must conduct a ***chronological analysis*** for each topic. That is, one dimension must be the time horizon from the distant past to the recent past. For example, the introduction of motor bikes in late 2020 and the COVID pandemic must have impacted the customers' demand for bike sharing. Their impacts can only be seen from a chronological analysis. (ii) You must utilize the weather and holiday data in your study. You do not need to use both in each analysis topic.

But ultimately, each of them must be used in some of your analysis topics. Utilizing the regional data is optional; but doing so may help you receive a higher mark to reward your additional effort.

You are expected to use SAS Enterprise Guide (EG) for this project. To begin with the ETL (extract, transform and loading) process, you need to prepare your data in proper tables that will go into SAS. That is, you need to create tables in the SAS environment.

Whenever you want to conduct an analysis, you must write a query to select relevant attributes by properly joining multiple tables to obtain a resultant table for specific analysis. See Appendix for using some common data analysis and visualization functions of SAS EG. More features of SAS EG will be introduced in a tutorial session later. (Note: it is possible that you may not be able to plot your desired graphs using SAS EG. If necessary, you may use other software such as Excel for graphing.)

Finally, please note that the management (or the LIC) does not know anything beyond this project statement. Therefore, you need to use your own judgement and make necessary and reasonable assumptions when doing this project. Make sure to present all assumptions made in the project.

Project Deliverable

Your group will submit a **PowerPoint file with your video, audio narration recorded, and speaker notes.** You should write your speaker notes in the Notes Pane for each slide. When you are recording your presentation video, you will speak following your own speaker notes in each slide. This will enable the LIC to both listen to your narration AND read your speaker notes when marking your project. While it is preferred that you turn on camera to show your face when you are making the presentation, it is also known that Mac users may not be able to show their faces on the PPT.

In addition to your PPT submission, you also need to submit an appendix (in pdf format) that contains supporting materials and queries. You should provide a good referencing in your appendix such as "this query supports table x or figure y on slide z".

DO NOT TURN IN A VIDEO FILE. PowerPoint includes a feature for recording slides. Here is a step-by-step reference:

https://www.ou.edu/cas-online/website/documents/Narrated%20Powerpoint%20(Office%20365).pdf

Follow the steps for "Preparing to Record" and "Recording Narration." You should ignore the last paragraph of this document on P. 4 and do not convert the PowerPoint file to a video file.

Your presentation should be limited to **8 minutes** with **no more than 10 static** slides that contain no animations or 'movement' of any description. In each slide, please properly place your video box so that it does not cover any important content.

Slide structure

The following is the required structure of your PPT presentation:

- 1st slide: Introducing group members, including your group name (column E on the group signup spreadsheet, e.g., H16A Group 2).
- 2nd slide: Your 2nd slide must display the following table only, which should be filled and contain summary information of your analysis topics. The following is merely an example:

No.	Topic Description	Chronological Analysis	Weather Analysis	Holiday Analysis	Regional Analysis	Note (e.g., special efforts that you want to the marker to know)
1	We study the weekly demand pattern and how it changes over the past 10 years, under different weather conditions.	√	√			
2	We study how different holidays influence the demand over the past 10 years	√		√		
3	N/A					

  - Column 1, No.: no more than 3 topics should be presented.
  - Column 2, Topic Description: briefly describe what you do in this topic
  - Columns 3 – 6: tick if the corresponding analysis is involved in your topic
  - Column 7, Note: If you have anything that you want the marker to know (e.g., special efforts), please write here.

- The next 1 - 2 slides: Briefly describe how you prepare the data for analysis, including how you clean data, manage missing information, and how you organize tables that go into SAS.
- The rest slides: use Analysis Topic I, Analysis Topic II, up to Analysis Topic III as the slide titles. For each topic, you should describe the description of the analysis, major findings (in terms of data visualization such as charts), business insights and recommendation.
- Please be aware that Moodle does not take a submission with file size larger than 200 MB. A 8-minute PPT with video will not automatically make your file large, but a fancy PPT theme, and/or using some original, high-definition images could easily make the file size exceeding 200MB. Please pay attention to this implicit limit on file size as well.

Marking guideline

Item (%)	Description
Data preparation (25%)	Do you properly manage missing data? Do you properly preprocess the tables used by SAS?
Quality of the data analysis (40%)	Are your analysis topics interesting and are not trivial? Are your analysis topics meaningful to the CaBi business? Have you properly analyzed the data with the right functions or steps?

	Have you provided proper data visualization (for example, table or graph) to present and support your analysis? Are there special efforts invested in processing or analyzing some data?
Quality of business insights obtained and recommendation (25%)	Do you obtain business insights from the data? Are your obtained insights helpful for business? Do you provide proper recommendations to make use of the obtained insights?
Presentation and recording quality (10%)	Is your presentation clear and effective to professional standards?
Total (100%)	

Appendix: Using Enterprise Guide for Data Analysis and Visualisation

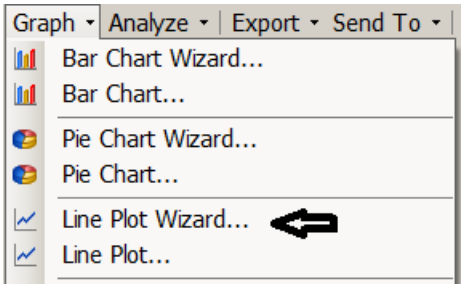
Given a data file opened in SAS Enterprise Guide, you can see some analysis and visualisation functions available (from the tool bar below).

Most functions are straightforward to use. Graphs can be found under Graph; some useful analysis tools can be found under Analyze in the tool bar. You are expected to try them by yourself.

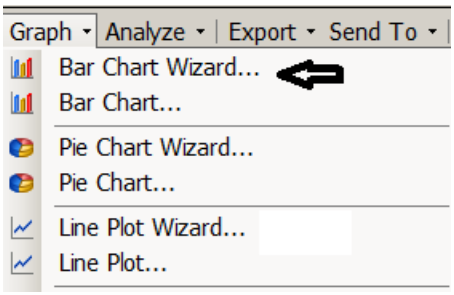
Note that the data visualisation functions only apply to a SAS data file only. When you write a query, before you can graph the table of the query outcome, you need to save the result table as a SAS data file using "create" statement, which has been introduced previously.

Graphing:

Line Chart



Bar Chart

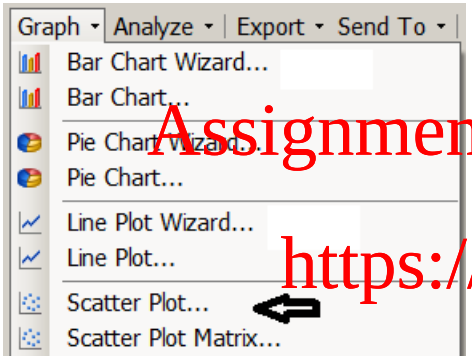


Histogram

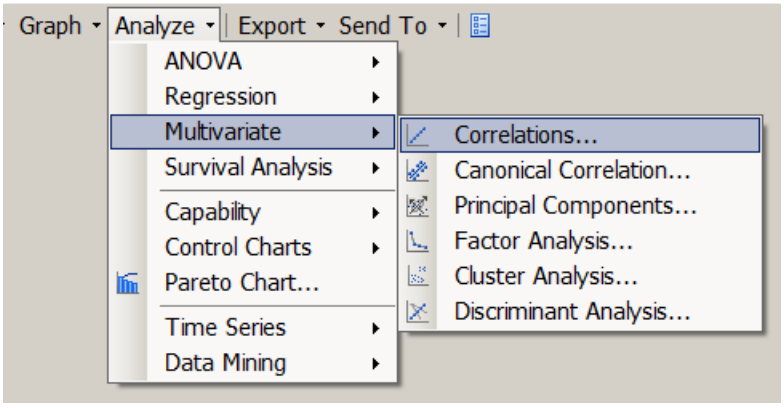
If you are not familiar with the concept of histogram, please read the following site about histogram. To plot a histogram, choose Bar Chart Wizard. In Step 2 out of 4, choose Percentage for the Bar height.

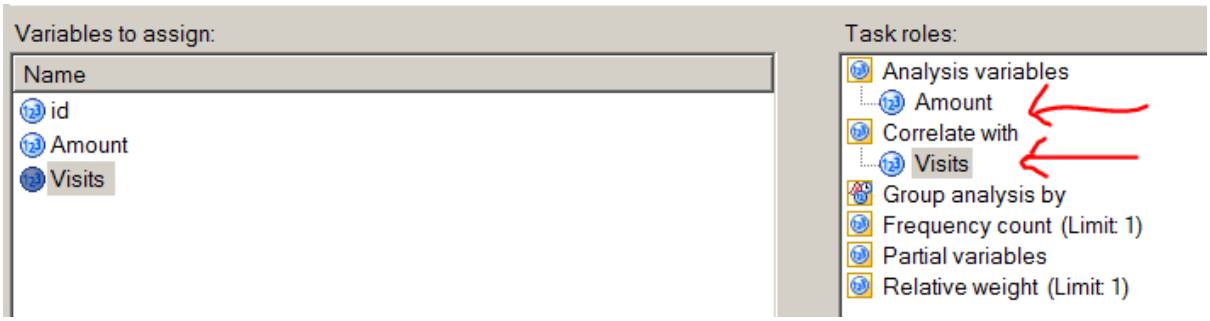
Correlation Analysis

You may plot a 2D scatter chart first for the two variables that you want to study their correlation.



If a correlation is revealed from the scatter chart, you may also calculate the exact correlation between these two variables. Assume these two variables are “Amount” and “Visits”. The following figures show how their correlation can be calculated.





Drag Amount and Visits from the left pane to the right pane.