BUSAN 300: DATA WRANGLING

PROJECT REPORT

DUE DATE: FRIDAY, 11TH JUNE @ 11:59PM

NAME: SHAKEEL ALI
UPI: SALI129

3.5 Task: Pose Questions

The three questions which I would like to be answered is how does Cost of Living, Homicide and Tourism look per region?

- In hindsight, I will be looking at:
  - The Average Cost of Living Index Per Region from 2016 to 2020.
  - Average Homicide Rate Per Region from 2016 to 2020
  - Total Number of Tourists Per Region from 2016 to 2020

3.6 .1 Project Summary Section

This project report details the necessarily measures taken to wrangle and combine datasets from various sources consisting of various formats in order to answer three questions. These questions on a holistic level were looking how Cost of Living, Homicide and Tourism look per regions? This was measured by answering sub-questions of finding:

- What the Average Cost of Living Index Per Region was from 2016 – 2020.
- Average Homicide Rate Per Region from 2016 to 2020.
- The Total Number of Tourists Per Region from 2016 to 2020.

Tableau – a data visualization tool was employed to graphically present the findings. It was concluded that the region of Australia and New Zealand, despite being having the highest cost of living index in the world had surprising the lowest average homicide rate. What made it more surprising was that despite the region of Australia and New Zealand being the safest region in the world – taking the homicide rate into account, it has the lowest accumulated tourist arrival from 2016 to 2020.

3.6.2 Wrangling Details Section

The initial datasets were sourced from various organizational websites. Although data for some of these topics were available and dated back to 1970s, I only used data pertaining to Years 2016 – 2020 in this project. The datasets as well as their format are given below:

- Corruption Index [Source: https://www.transparency.org/en/cpi/2020/index/nzl]
  - CPI2020_GlobalTablesTS_210125.xls [Data from 2012 – 2020]
  - CPI2020_SignificantChanges_210125.xls [Data from 2019 – 2020]
    - Only the dataset from the first file was used as it had years of data which I wanted to analyze i.e., 2016 – 2020.
    - There were 34 columns [Country, IS03, Region, CPI scores Year, Ranks Year, Sources Year, Standard Error Year. There were around 180 rows of data.
    - The rank for Year 2016 was missing so I had to calculate it based on its CPI score in another sheet.

- Cost of Living Index [Source: https://www.numbeo.com/cost-of-living/rankings.jsp]
  - The problem with this source was that their data was only viewable as a table format on their website. Hence, I had to copy and paste into excel sheet. I ensured to copy and paste each year [2016 – 2020] into different workbook to maintain integrity.
  - It contained 8 fields with 122 rows of data for each year!
  - I wrangled all of the attributes and the country column served as basis of VLOOKUP [Meaning that those other countries which weren't found here were not included in the combined datasets!]

- Education Index [Source: http://hdr.undp.org/en/indicators/103706]
  - Education Index.csv
    - Although I initially submitted this dataset in my proposal, I didn't use. This is because the three questions which I answered didn't need this dataset.
    - It had 60 columns with 212 rows of data

- Homicide Per Country [Source: https://data.worldbank.org/indicator/VC.IHR.PSRC.P5]
  - API_VC.IHR.PSRC.P5_DS2_en_xml_v2_2261319.xml
    - The data was in an XML format.

        - There were 4 fields [These fields were redundant as it had country name, year, etc. which was found in other datasets. NOTE: Region wasn't included in this dataset]. It also had 16288 rows of data which I wrangled!
        - Only the homicide rate was used in the combined datasets and analysis.
        - Years dated back to 1970s, but unfortunately even though they listed the year 2020, most of the countries values were null.
        - I changed the Null values into N/A.

- Human Freedom Index [Source: https://www.cato.org/human-freedom-index/2020]
    - human-freedom-index-2020.json
        - The data was in a JSON format.
        - There were significant number of attributes in this. However, I only wrangled 10 fields [Some of them were redundant but were needed for VLOOKUP] and had 1782 rows
        - Unfortunately, even though I did combine this into final dataset, I didn't use it to answer my question.

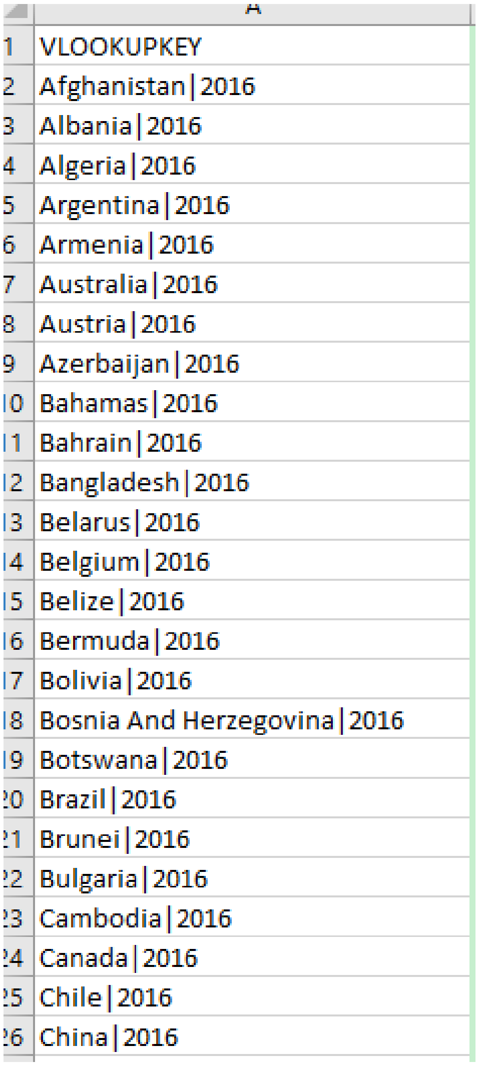
- International Tourism [Source: https://data.worldbank.org/indicator/ST.INT.ARVL]
    - API_ST.INT.ARVL_DS2_en_xml_v2_2252568.xml
        - The data was in XML format.
        - There were 4 fields which I wrangled; Key, Country, Year and Tourist No (Total Number of tourists who visited a particular country per year).
        - There were 16,287 rows of data!
        - The final combined dataset has only the Tourist No.

- Migration [Source: https://www.un.org/development/desa/pd/content/international-migrant-stock]
    - UN_MigrantStockTotal_2019.xlsx
        - Although I initially submitted this dataset in my proposal, I didn't use. This is because the three questions which I answered didn't need this dataset.
        - Furthermore, the datasets were only available in 5-year increment rather than for each year. Therefore, I decided not to use this dataset.
        - It had several sheets of data with may fields, and around 282 rows of data.

- World Happiness [Source: https://www.kaggle.com/mathurinache/world-happiness-report]
    - 2015 – 2020. csv
        - Each of these csv files had the same number of attributes i.e., 13 and 157 rows of data.

Steps Performed to Combine the Data

- The initial dataset I used was of Cost of Living. This meant, that the VLOOKUP performed later to merge data-sets looked at the Look-up value of country found in the Cost-of-Living data-set.
- I also created a VLOOKUP key by combing Country&"|"&Year as some of the future datasets needed two conditions in order to be wrangled successfully.



	A
1	VLOOKUPKEY
2	Afghanistan|2016
3	Albania|2016
4	Algeria|2016
5	Argentina|2016
6	Armenia|2016
7	Australia|2016
8	Austria|2016
9	Azerbaijan|2016
10	Bahamas|2016
11	Bahrain|2016
12	Bangladesh|2016
13	Belarus|2016
14	Belgium|2016
15	Belize|2016
16	Bermuda|2016
17	Bolivia|2016
18	Bosnia And Herzegovina|2016
19	Botswana|2016
20	Brazil|2016
21	Brunei|2016
22	Bulgaria|2016
23	Cambodia|2016
24	Canada|2016
25	Chile|2016
26	China|2016

- I then combined “Corruption Perception Index” and “Corruption Rank (Lower Is Better)” using VLOOKUP and setting the last parameter to “FALSE” in order to not get approximate match.
- Similarly, I used VLOOKUP to get “Happiness Rank”, “Happiness Score” and “Region” from the World Happiness csv files for each year.
- Moving on to Human Freedom Index.
  - This dataset was in a JSON format.
  - The attributes I got were “Economic Freedom Score”, “Economic Freedom Rank”, “Personal Freedom Rank”, “Human Freedom Score” and “Human Freedom Rank”
  - The thing with this specific JSON file was that it listed all its values for the first field in a row-format, then all its values for its second field in a row-format.
  - Therefore, I first opened it in VScode, to format the document properly.
  - I then cut and pasted it into Atom as I like to do RegEx using Atom.
    - I then searched for the attributes I was interested in. The reason why I did this was because this dataset had lots of fields! It was easier to just copy and paste data for the attributes I needed into separate tabs of Atom [To have integrity]
      - Year
      - ISO_Code
      - Country
      - pf_score (Personal Freedom Score)
      - pf_rank (Personal Freedom Rank)
      - ef_score (Economic Freedom Score)
      - ef_rank (Economic Freedom Rank)
      - hf_score (Human Freedom Score)
      - hf_rank (Human Freedom Rank)
  - I then performed these steps which are detailed here:
    - **Link to steps performed:** https://drive.google.com/file/d/1J87cAhixKDMDstx2JzIr_JkvMliIPR1r/view?usp=sharing
    - **Link to Freedom dataset wrangled in xlsv:** https://drive.google.com/file/d/1HWDMd6uR22pPhQ6mRurJc6EyksWihhHu/view?usp=sharing

- Moving on to Homicide Dataset wrangling.
  - As mentioned, this dataset was in XML format.
  - I was only interested in Key, Country or Area, Year and Homicide Value.
    - The first thing I did was open it via VScode, format it nicely and then copy and paste into Atom to perform RegEx queries. [I prefer Atom over VScode but one can do it in VScode as well]

    - The steps of how I converted the XML into CSV and then imported in Excel using Power Query is detailed here:
      - **Link to steps performed:** https://drive.google.com/file/d/1lyxIEvYkcCzKEg7qtQrG2IHgtat0vugQ/view?usp=sharing
      - **Link to homicide wrangled data in txt, csv and xlsx:** https://drive.google.com/drive/folders/1F_5I9XAyxnDHsIWsVyMJyQh_Ndk1A63b?usp=sharing
- Lastly, I used VLOOKUP joining Country&”|”&Year

  - Moving on to International Tourism dataset wrangling.
    - As mentioned, this dataset was in XML format.
    - I was only interested in Key, Country or Area, Year and Tourist Number
    - The first thing I did was open it via VScode, format it nicely and then copy and paste it into Atom to perform RegEx queries. [I prefer Atom over VScode but one can do it in VScode as well]
      - The steps of how I converted the XML into CSV and then imported in Excel using Power Query is detailed here:
        - **Link to steps performed:** https://drive.google.com/file/d/1PN3F_c8cT-ajOtN65NQPk8lQk5Z1wSkv/view?usp=sharing
        - **Link to tourism wrangled data in txt, csv and xlsx:** https://drive.google.com/drive/folders/1v7fJ7SwULThEv0s-gResg8xTf6NFUx9W?usp=sharing
  - Lastly, I used VLOOKUP joining Country&” |” &Year
  - Below is as link which has a “VLOOKUP” excel file whereby all the datasets got transferred from all the reference datasets of Tourism, Homicide, and Freedom. https://drive.google.com/drive/folders/1tThFVJtaYCcnF6eujKXemlF6ieznJixZ?usp=sharing

- I then copy and pasted it into a new workbook so that the VLOOKUP cells don't give #REF error. The final Excel workbook can be found with the below link together with the final meta data [It explains the color coding, etc.]

  Link to Final Combined Datasets and Metadata of final data sets:

  - https://drive.google.com/drive/folders/1W7HZZhgK8StnIuCWZDS7jCY_3F19zhzf?usp=sharing

  Link to BUSAN 300 Project folder which contains everything:

  https://drive.google.com/drive/folders/1jae44LTbJUyNLa6-wV08MAjdRDGwbqNo?usp=sharing

3.6.3 Question Answers Section

After the datasets were combined using various means such as VLOOKUP, RegEx and Power Query, I used Tableau in order to go some visualization in order to answer my questions. The Excel workbook was uploaded into the Tableau desktop app whereby which aided me to prepare a data-viz. The question which I wanted to be answered was how does Cost of Living, Homicide and Tourism look per region?

- In hindsight, I looked at:
  - The Average Cost of Living Index Per Region from 2016 to 2020.
    - The region of New and Australia has the highest cost of living in the world with an average Cost of Living Index of 77.65, barely inching Western Europe which has an index of 76.65. In comparison, Southern Asia has the lowest average cost of living index of 28.87. This means that Western Europe and Australia and New Zealand has more than 2 times cost of living index of Southern Asia. This analysis was done in the form of a Tree Map.
  - Average Homicide Rate Per Region from 2016 to 2020
    - Region of Latin America and Caribbean has the highest average homicide rate in the world of 20.06 killed per 100,000 people. In comparison, region

    of Australia and New Zealand – despite being the region with the highest average cost of living has one of the lowest homicide rates in the world with a mere homicide rate of 0.88 per 100,000 people. This analysis was done in the form of a Map.

  - Total Number of Tourists Per Region from 2016 to 2020
    - The region which has had the most accumulated tourists from 2016 – 2020 is Western Europe with over a 2 billion visiting their shores – almost doubling the Central and Eastern European region which is the second most visited region, and almost certainly doubling the North American region which has the third most visited region. In comparison, Australia and New Zealand has the lowest number of tourist arrival. This analysis was done in the form of a Tree Map.

NOTE: The visualization can be accessed by the link provided below on Tableau Public. You can also download the Tableau workbook which details all the steps I took.

https://public.tableau.com/app/profile/shakeel.ali3390/viz/BUSAN300Project/Dashboard2