Term End Project Milestone-1

Prashant Raghuwanshi

DSC, Bellevue University

DSC540-T301 Data Preparation (2221-1)

Guided by Professor [Catherine](mailto:Catherine.williams@bellevue.edu) Williams

09/26/2021

ABSTRACT:

The first milestone of Data Preparation project will be to select the data we want to work with. we will need to select 3 different data sources that have different file types of information – and the data will need to have a relationship between them. If one doesn’t exist, you will have to create one. It is likely we will need to manipulate the data to create a relationship. Finding the data, you want to work with for this project, will likely be the hardest part of the project.

Term End project Milestone -1

1. **INTRODUCTION**

Please find the below requirements for data preparation project requirement for milestone -1 for for the term-end project.

* 3 data sources, along with a description of each one (links to each are fine, no need to submit the actual data)
* The relationships between them, or the relationship you will make between them
* What you believe you will have to do to the data to accomplish all 5 milestones and what your interpretation is of what the data means (you could provide a data dictionary or a summary of what the data is) – should be at least 250 word

1. **Data Source 1**

Data for Airbnb in Amsterdam

Inside Airbnb

Inside Airbnb provides data compiled from the Airbnb web-site for listings available for Amsterdam.

According to Inside Airbnb data for Amsterdam, compiled on December 2017, there are:

6,183 "Entire homes/apartments" (33% against the total number of listings) that were estimated to be booked for more 60 nights a year (and against the law)

1,299 listings that are available to accommodate more than 4 people (and against the law - this is underestimated, because it doesn't account for multiple listings in the same apartment);

A conservative occupancy model has been built in order to estimate Occupancy Rates, Income per Month and Nights per Year.

Data Source link:

<http://insideairbnb.com/get-the-data.html>

File Details :

* listings.csv, contains the list of all airbnb house listing in Amsterdam.
  + - About this file : Total Columns 74
    - Total Records : 2481
* reviews.csv This data file conatins the review for listed house in Amsterdam
  + - About this file : Total Columns 6
    - Total Records : 203701
* calendar.csv, This data contaion the varying price list of listed house for next 4 months
  + - About this file : Total Columns 7
    - Total Records : 905201

**The relationships key** between all of them : Listing ID

This data helps in predicting the arbnb price trends for next up coming years

1. **Data Source 2**

[Employee\_work\_preference](https://gssdataexplorer.norc.org/projects/106024/extracts/80310)

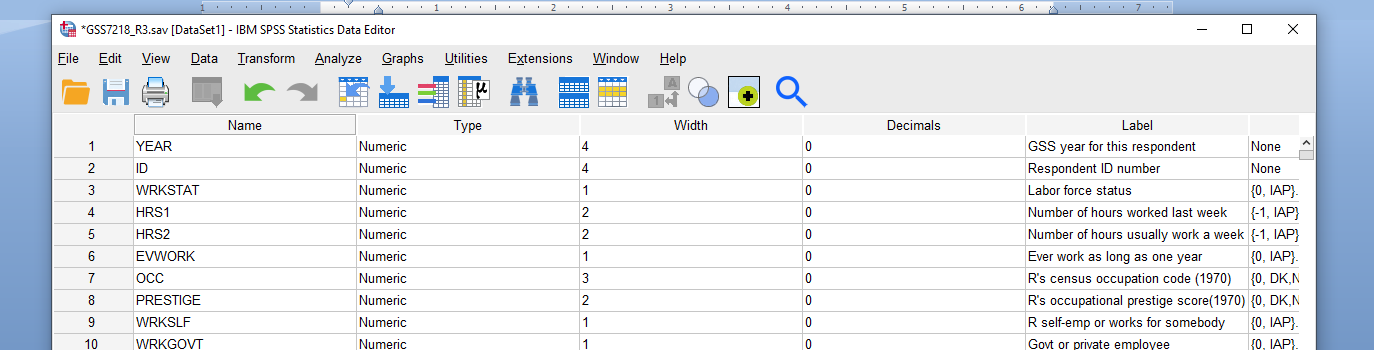
This data contains different survey data related to employee work preference

Data Source link:

<https://gssdataexplorer.norc.org/projects/106024/extracts>

File Details :

* Employee\_work\_preference.xls.
  + - About this file : Total Columns 11
    - Total Records : <10000
  + **Variables:**
  + YEAR(Gss year for this respondent )
  + EXTR2001(Did r receive perf based payment in 2001)
  + COMPPERF(Size of perf based pay depend on profits)
  + EXTRAPAY(Eligible for performance based pay)
  + MARITAL(Marital status)
  + OCC80(Rs census occupation code (1980))
  + INDUSTRY(Rs industry code (1970))
  + OCC(Rs census occupation code (1970))
  + WRKSTAT(Labor force status)
  + ID\_(Respondent id number)
  + EXTRAVAL(Total dollar value of pay in previous year)
  + **Case selection:**
  + (Combined.year >= 2008 AND Combined.year <= 2018)
  + **Years Included:**
  + 1972-2018
  + **Formats:**
  + Excel Workbook (data + metadata)
  + Excel Workbook (metadata only)
* GSS7218\_R3.sav This data file conatins codes for OCC
  + - About this file : Total Columns <6000
    - Total Records : <60000



**The relationships key between all of them** : OCC

This data helps in predicting the employee work preference trands based on compensation for next up coming years

1. **Data Source 3**

AirStats: Data on airports around the world

Data on airports around the world, their codes, official websites, and much more

Data Source link:

<https://www.kaggle.com/patrasaurabh/airstats-data-on-airports-around-the-world?select=airports.csv>

File Details :

1. "airports.csv" contains the list of all airport codes. Some of the columns contain attributes identifying airport locations, other codes (IATA, local if exist) that are relevant to identification of an airport.

**About this file : Total Columns 18**

**Total Records : 67313**

2) Countries.csv This data is required to interpret the country codes in airports.csv data

**About this file : Total Columns 6**

**Total Records : 248**

1. Region.csv This data is required to interpret the regional codes in airports.csv

**About this file : Total Columns 8**

**Total Records : 3964**

**The relationships between them : Region ID , Country ID** with airport.csv

This Data is going to use by fleet management companys to estimate the business apportunity for airport fleet service operations across the world