Name: Astrid Fuentes
Date: February 22nd, 2021
Title: Final Project – Part I

1. Dataset 1: Environmental Radiation Surveillance Gamma Radiation Readings: Beginning 1995

https://healthdata.gov/dataset/environmental-radiation-surveillance-gamma-radiation-readings-beginning-1995/resource-3#{view-graph:{graphOptions:{hooks:{processOffset:{},bindEvents:{}}}},graphOptions:{hooks:{processOffset:{},bindEvents:{}}},view-map:{geomField:!Location}}

1.1 Introduction

This dataset provides gamma radiation readings collected through Environmental Thermoluminescent Dosimeter (TLDs) devices placed at various facilities and locations in New York State. The Environmental TLDs provide a quantitative measurement of the radiation levels in the area in which they are placed. This dataset can be used by the general public, researchers and facility staff to evaluate environmental radiation levels at various locations. Facility operators may use data for inter-comparison and trending purposes. The facility operator performs this type of monitoring to meet the requirements of its licensing agency (e.g., a federal agency or the New York State Department of Environmental Conservation). The data set contains data from 1995 to 2018 in quarters.

1.2 Research questions

- Has gamma radiation been increasing over the years in different facilities in New York State?
- What is the predicted 2021 and 2022 gamma radiation level in New York state?

1.3 Approach

I would like to start by importing the data set in R studio and doing some preliminary descriptive analytics as well as some cleaning of my dataset. Then, I will evaluate if my data is normally distributed and do some histograms, scatter plots, etc to look for trends and outliers. I would measure correlation between variables in order to determine the best variables to keep in order to perform regression. I will try with linear regression first and take it from there. Source data is health.data.ny.gov.

1.4 How your approach addresses (fully or partially) the problem?

The first question can be fully answered using some plots. The second question, I will be able to answer it if I can find a model that seems like a good fit to predict gamma radiation levels in New York State.

1.5 Data

The data set contains 1912 records of quarterly gamma radiation measurements from 1995 to 2018. The fields contained in the data set include: Facility/Operator, Location, Year, Quarter, Gamma radiation, Reported Unit, Comment. I plan on using all the variables except maybe the comments. As part of my data cleaning steps, I will make sure all different Facilities, locations, and reported units and consistent and if they are not, I will make sure I make them consistent across the data set by applying casing and other matching functions. I will also make sure all reported values for gamma radiation are converted to the same unit If needed. I will get rid of rows that did not report any gamma radiation.

1.6 Required Packages

I will need to use ggplot2 for my graphs, pastecs for stats descriptions, ggm for correlations, and maybe other packages that I will not be able to identify at this time but when I actually need them.

1.7 Plots and tables

I plan on using scatter plots, histograms, and regression plots. I will probably use a correlation table and pcor().

1.8 Questions for future steps

- Since the data is presented in quarters, it is a good candidate for time series analysis. This is probably a little bit more advance and would be something interested to consider in future steps.
- I would be interested in analyzing similar data from other states.

2. Data set 2: Disney Movies

https://www.kaggle.com/prateekmaj21/disney-movies

2.1 Introduction

This dataset provides information about Disney movies, including title and revenue. I would be interesting to study this data just for fun.

2.2 Research questions

- Is there a particular genre within Disney movies that has been more successful?
- Is Disney's movies popularity based on revenue increasing or decreasing over time?

2.3 Approach

I would like to start by importing the data set in R studio and doing some preliminary descriptive analytics as well as some cleaning of my dataset. Then, I will evaluate if my data is normally distributed and do some histograms, scatter plots, etc to look for trends and outliers.

2.4 How your approach addresses (fully or partially) the problem?

I think my approach based on the data set is going to give a partial view of the problem. There are many other factors that could affect revenue in movies that might not be represented by this data set.

2.5 Data

The dataset contains 6 columns: movie_title, release_date, genre, rating, total_gross_income, inflation_adjusted_income.

2.6 Required Packages

I will need to use ggplot2 for my graphs, pastecs for stats descriptions, ggm for correlations, and maybe other packages that I will not be able to identify at this time but when I actually need them.

2.7 Plots and tables

I plan on using scatter plots, histograms, and line plots. I will probably use a correlation table and pcor().

2.8 Questions for future steps

- I would like to see the inflation adjusted cost of making each of these movies. It would be interesting to join that data to this data set and be able to calculate net revenue.
- I might come up with more questions as I see some of the descriptive statistics but at this time I do not have more.

3. Data set 3: NYPD Motor Vehicle Collisions

https://www.kaggle.com/new-york-city/nypd-motor-vehicle-collisions?select=MVCollisionsDataDictionary_20190813_ERD.xlsx

3.1 Introduction

This dataset provides information about motor vehicle collisions reported by the NYPD. The Motor Vehicle Collisions crash table contains details on the crash event. Each row represents a crash event. The Motor Vehicle Collisions data tables contain information from all police reported motor vehicle collisions in NYC. The police report (MV104-AN) is required to be filled out for collisions where someone is injured or killed, or where there is at least $1000 worth of damage.

3.2 Research questions

- I would like to find trends in this data that would be able to tell us if there is a correlation between sex and the number of accidents in NY City.
- Is there a specific location in NY City that has more accidents than the rest?
- Is there a particular vehicle make and/or model that is significantly involved in more accidents than the rest?
- Is there a particular contributing factor that causes more accidents in NY City?

3.3 Approach

I would like to start by importing the data set in R studio and doing some cleaning. Since the file has a lot of data, it would be easier to drop some columns and only keep those that I truly need. Some preliminary descriptive analytics will be needed as well as some cleaning to treat any possible typos, missing data, etc. Then, I will do some histograms, scatter plots, etc to look for trends and outliers. I would build hypothesis tests and see if I find statistically significant results. This is a dataset hosted by the City of New York. The city has an open data platform found here and they update their information according the amount of data that is brought in.

3.4 How your approach addresses (fully or partially) the problem?

I think my approach fully addresses the problem. Given that the data set has a lot of data, I will be able to obtain a lot of information from it.

3.5 Data

The dataset contains 1,612,181 records and the following 25 columns:

Column Name	Column Description
UNIQUE_ID	Unique record code generated by system
COLLISION_ID	Unique crash identification code
ACCIDENT_DATE	Occurrence date of collision
ACCIDENT_TIME	Occurrence time of collision
VEHICLE_ID	Vehicle identification code assigned by system
STATE_REGISTRATION	State where driver license was issued
VEHICLE_TYPE	Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other)
VEHICLE_MAKE	Vehicle make
VEHICLE_MODEL	Vehicle model
VEHICLE_YEAR	Year the vehicle was manufactured
TRAVEL_DIRECTION	Direction vehicle was traveling
VEHICLE_OCCUPANTS	Number of vehicle occupants
DRIVER_SEX	Gender of driver
DRIVER_LICENSE_STATUS	License, permit, unlicensed
DRIVER_LICENSE_JURISDICTION	NYPD, Port Authority, TBTA, MTA, etc.
PRE_ACDNT_ACTION	Going straight, making right turn, passing, backing, etc.
POINT_OF_IMPACT	Location on the vehicle of the initial point of impact (i.e. driver side, passenger side rear, etc.)

VEHICLE_DAMAGE	Location on the vehicle where most of the damage occurred
VEHICLE_DAMAGE_1	Additional damage locations on the vehicle
VEHICLE_DAMAGE_2	Additional damage locations on the vehicle
VEHICLE_DAMAGE_3	Additional damage locations on the vehicle
PUBLIC_PROPERTY_DAMAGE	Public property damaged (Yes or No)
PUBLIC_PROPERTY_DAMAGE_TYPE	Type of public property damaged (ex. Sign, fence, light post, etc.)
CONTRIBUTING_FACTOR_1	Factors contributing to the collision for designated vehicle
CONTRIBUTING_FACTOR_2	Factors contributing to the collision for designated vehicle

3.6 Required Packages

I will need to use ggplot2 for my graphs, pastecs for stats descriptions, ggm for correlations, and maybe other packages that I will not be able to identify at this time but when I actually need them.

3.7 Plots and tables

I plan on using scatter plots, histograms, and line plots. I will probably use a correlation table and pcor().

3.8 Questions for future steps

- Since the data set has many columns, I might not be able to study them all, some additional correlations and statistics could be calculated in future steps.
- I might come up with more questions as I see some of the descriptive statistics but at this time I do not have more.