Exploratory Data Analysis

Data Science Persistency of a Drug Project

DS HealthCare Group Members:

- Yuchi Chen
- Bolutife Akinlawon
- Alexis Collier
- Han-Fu Lin
- Runtian Wang

Agenda

- ➢ Problem Understanding
- ➢ Business Understanding
- ➢ Data Preparation
- ➢ Exploratory Data Analysis
- ➢ Conclusion

Problem Understanding

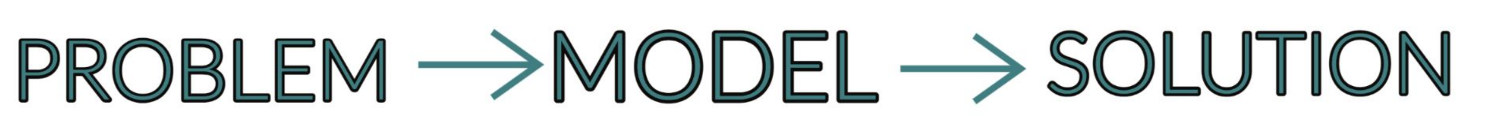
ABC Pharma is a pharmaceutical company that aims to automate the identification of the persistence of a drug.

In this data analysis, the various factors of a patient are evaluated to determine if they affect persistency. The data is collected for analysis to determine durability of the drug.

At the end of this project, we will suggest a model follow for deployment.

BUSINESS UNDERSTANDING

In summary, the task can be represented as follows:



PROBLEM → MODEL → SOLUTION

DATA PREPARATION

- Python was utilized for data preparation, as well as pandas library specifically.
- Data was cleaned and prepared for analyzation.
  - Method of Approach:
    - Look for null or missing values.
    - Identify values that are improbable.
    - Create visualizations of the data.

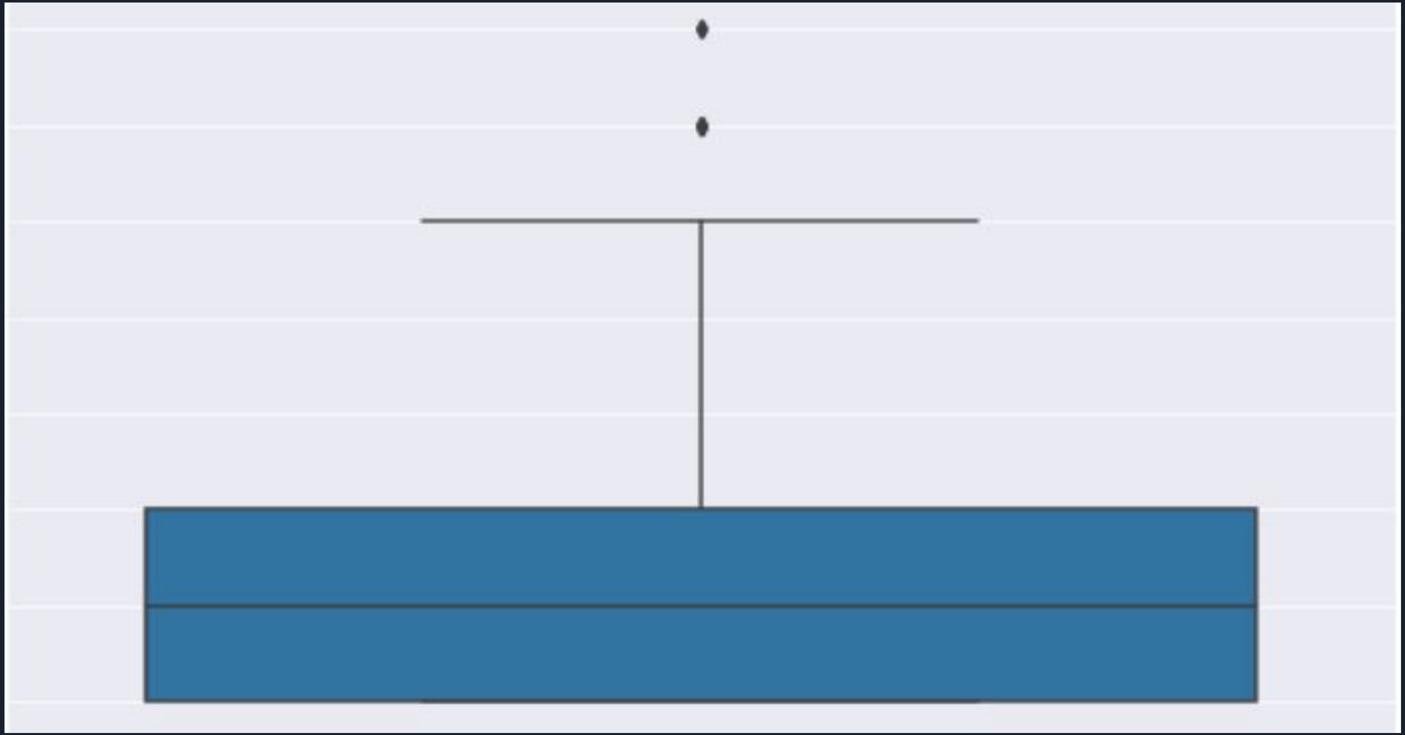
EXPLORATORY DATA ANALYSIS

- ➢ We analyzed the demographic data characteristics include gender, race, ethnicity, region, age bracket, and iodine indicator.
- ➢ The data variables are represented via stacked bars and correlation heatmaps.

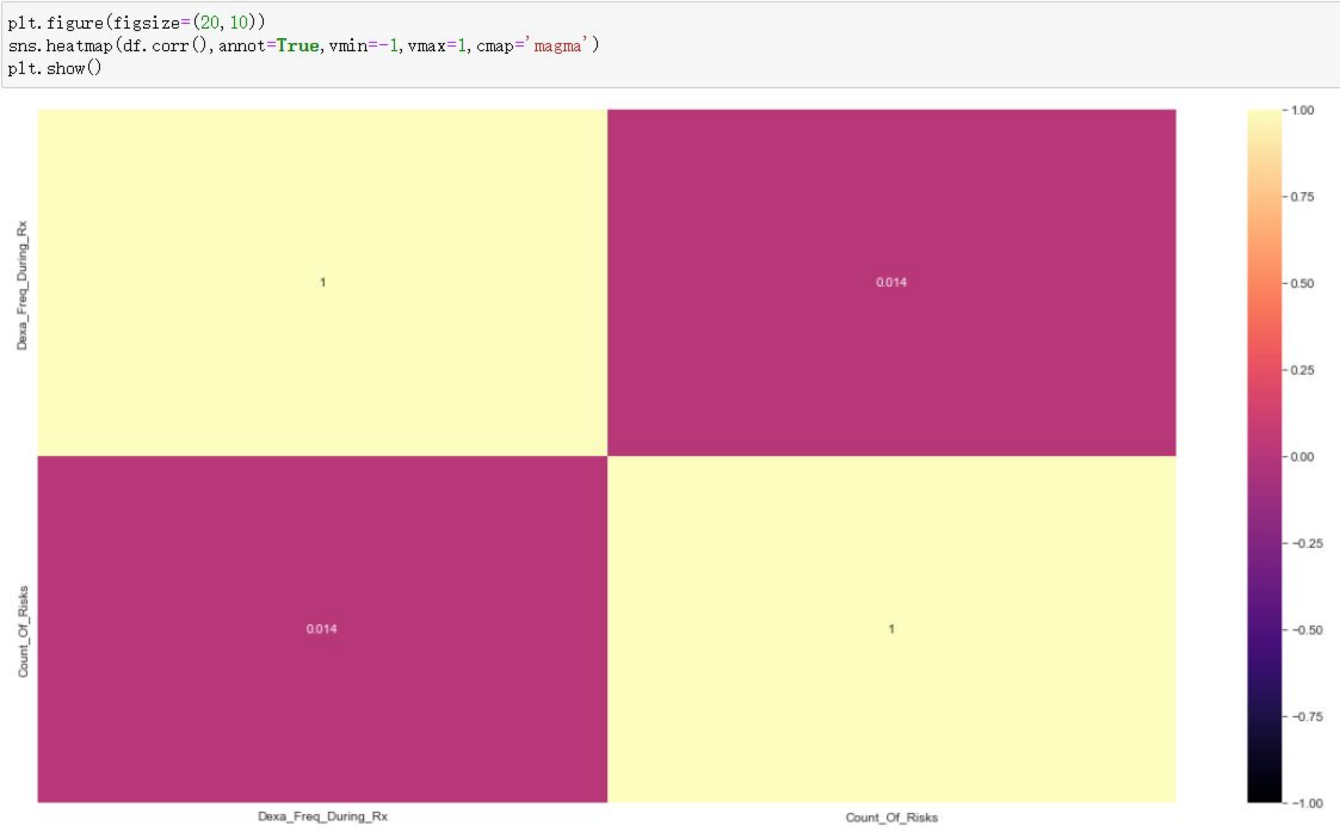
EXPLORATORY DATA ANALYSIS

- Some demographic observations include the following:
  - More significant number of female participants.
  - There are more persistent than non-persistent participants, indicating an imbalanced dataset.
  - The primary racial demographic of the dataset is caucasian (non-hispanic).
  - Most of the participants are 75 and older.
  - ‘Below 55’ age bracket is the least represented age demographic.

Outlier Test

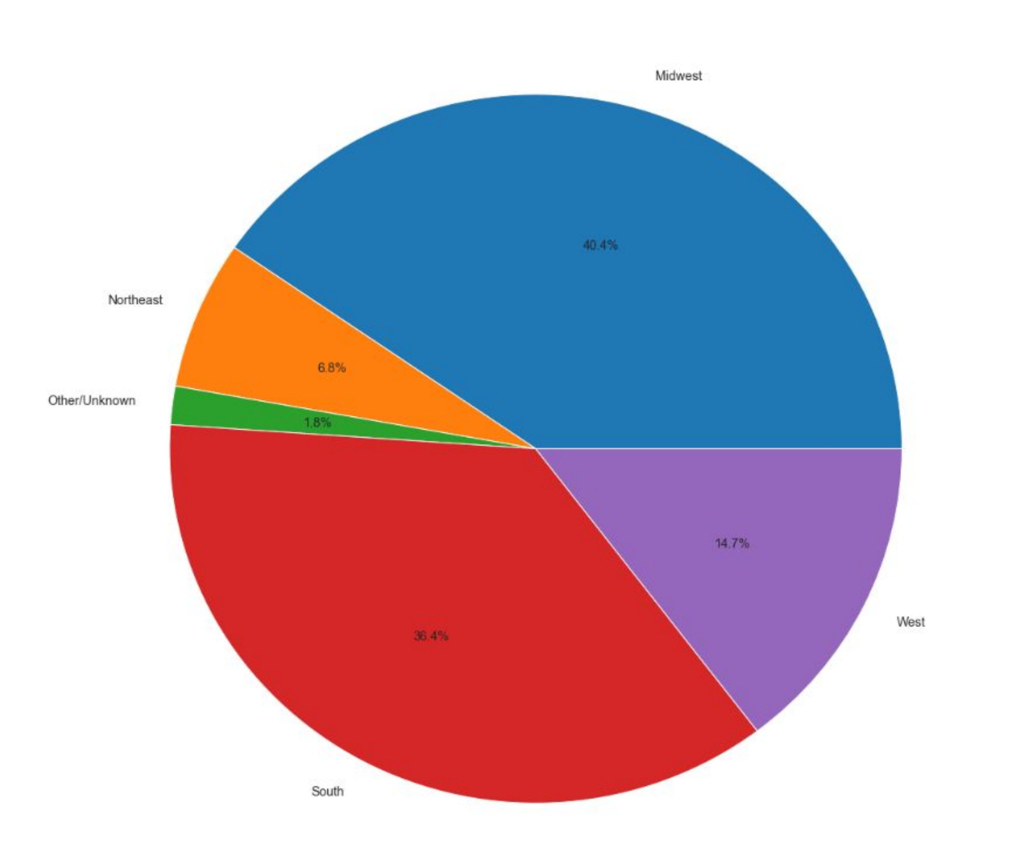


Heatmap

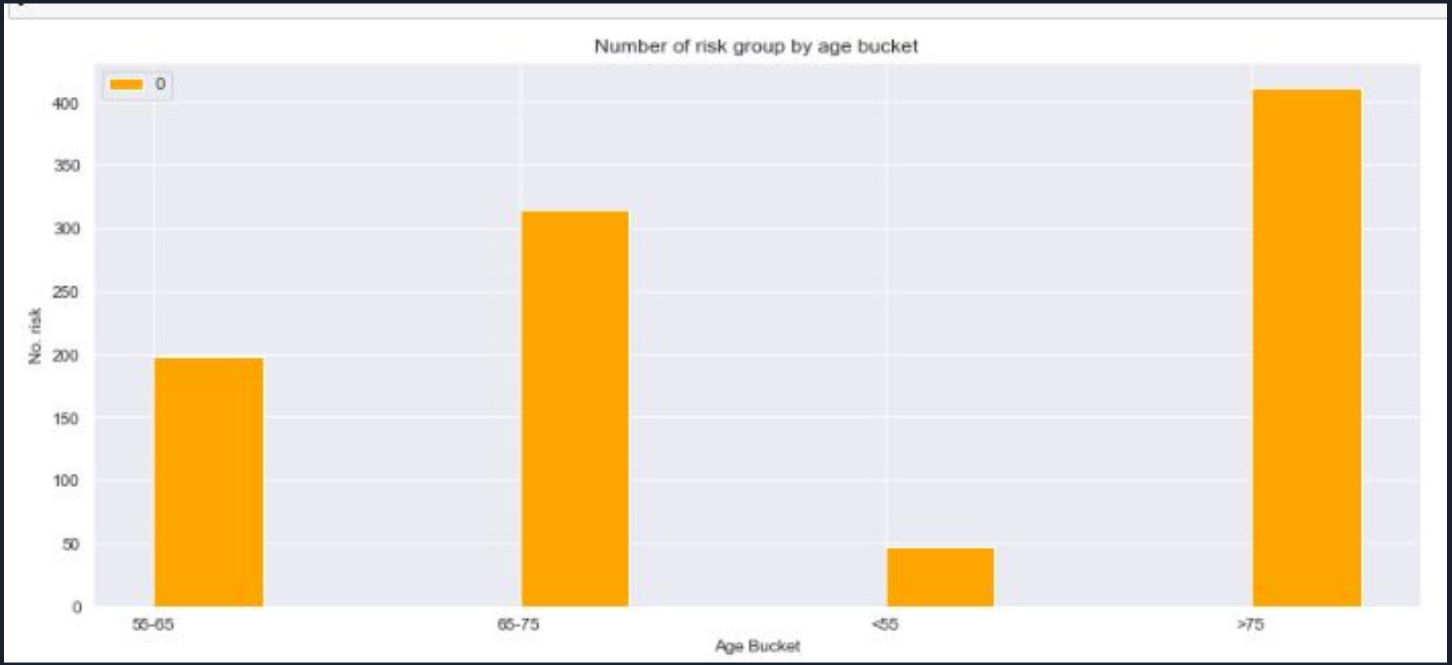


```
plt.figure(figsize=(20,10))
sns.heatmap(df.corr(),annot=True,vmin=-1,vmax=1,cmap='magma')
plt.show()
```

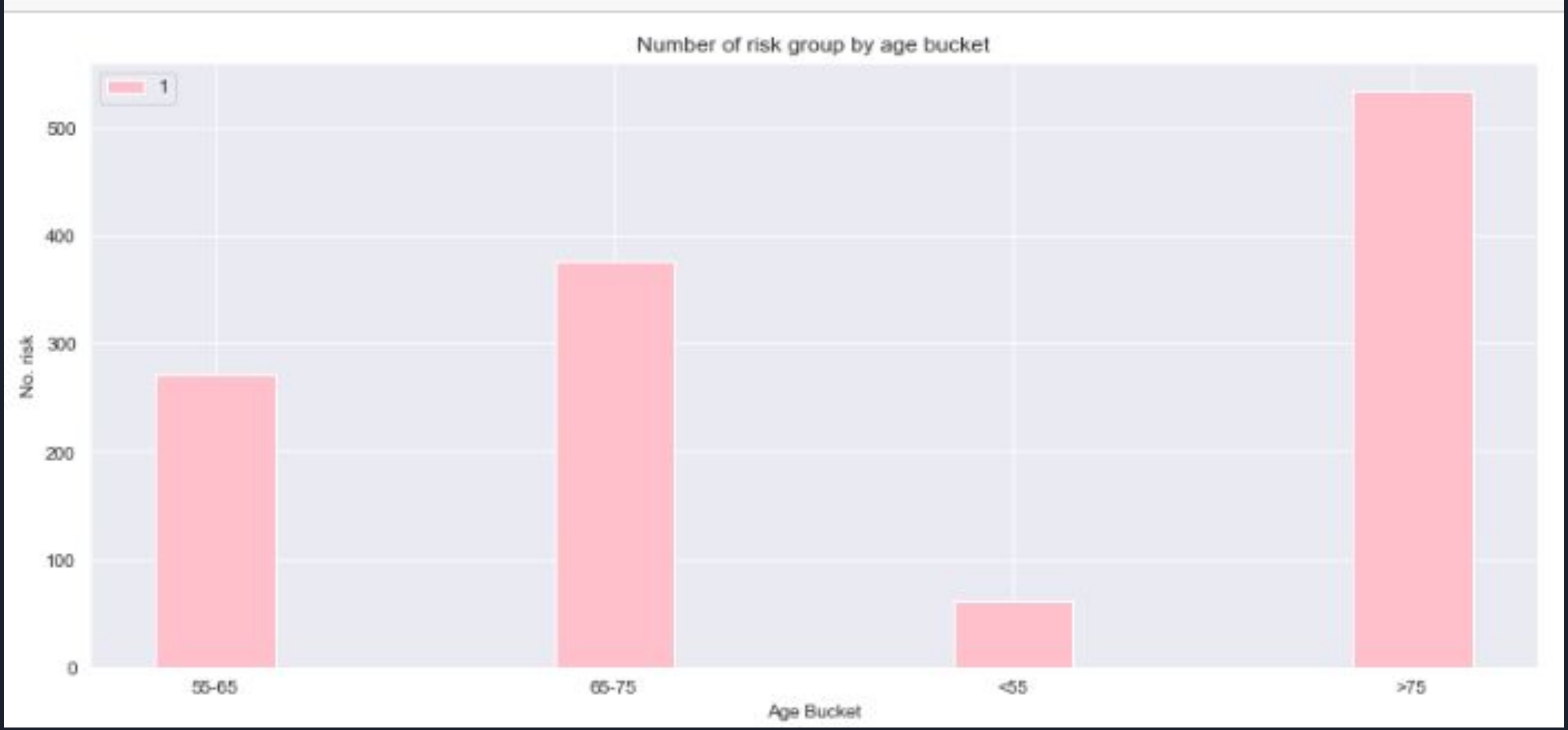
Data Amount by Region



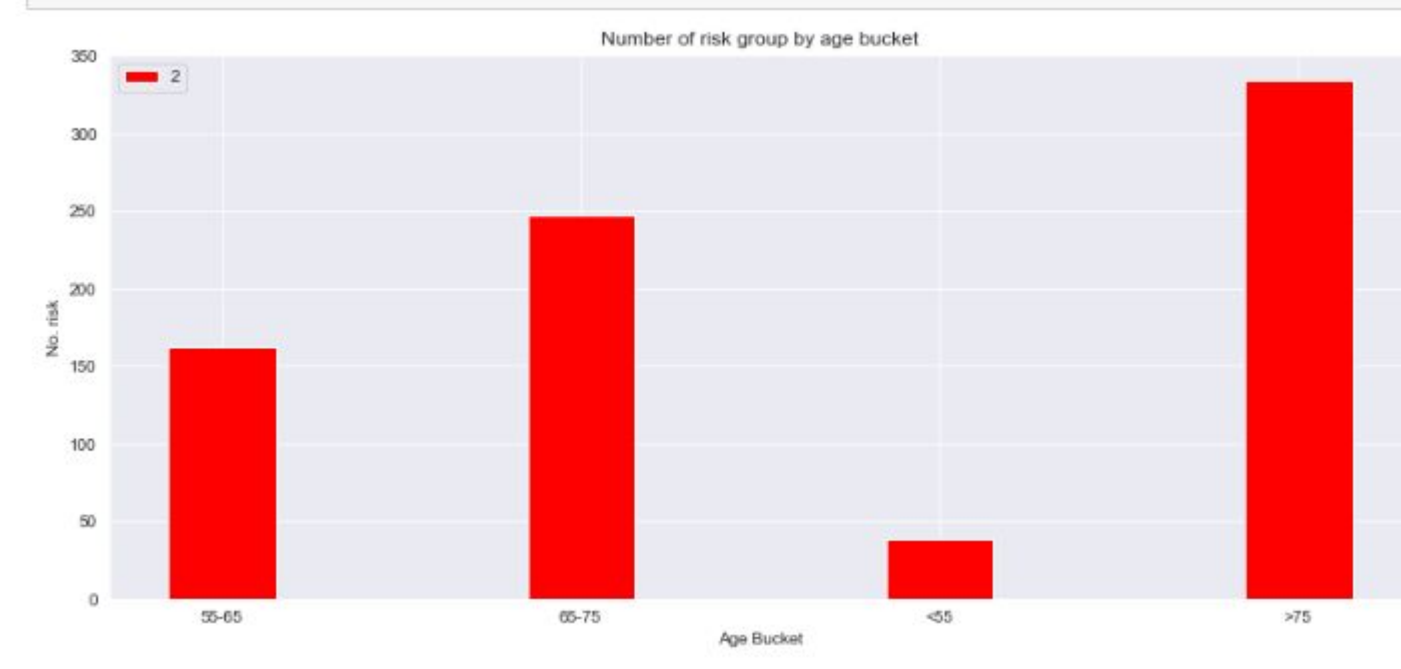
Risk Count by Different Age Group



Number of Risk Group by Age Bucket



As a result, the count of risk doesn’t show much difference in different age group.



Conclusion

Of the many factors that may affect a patient’s persistence to a drug is having risks, being of a certain race, ethnicity, age group, and the specialty of the HCP. Now that the factors have been identified, we need to identify a model that can create an automated process to predict whether a drug will be persistent.

The random forest classifier model would be the best used in this scenario.

Thank You