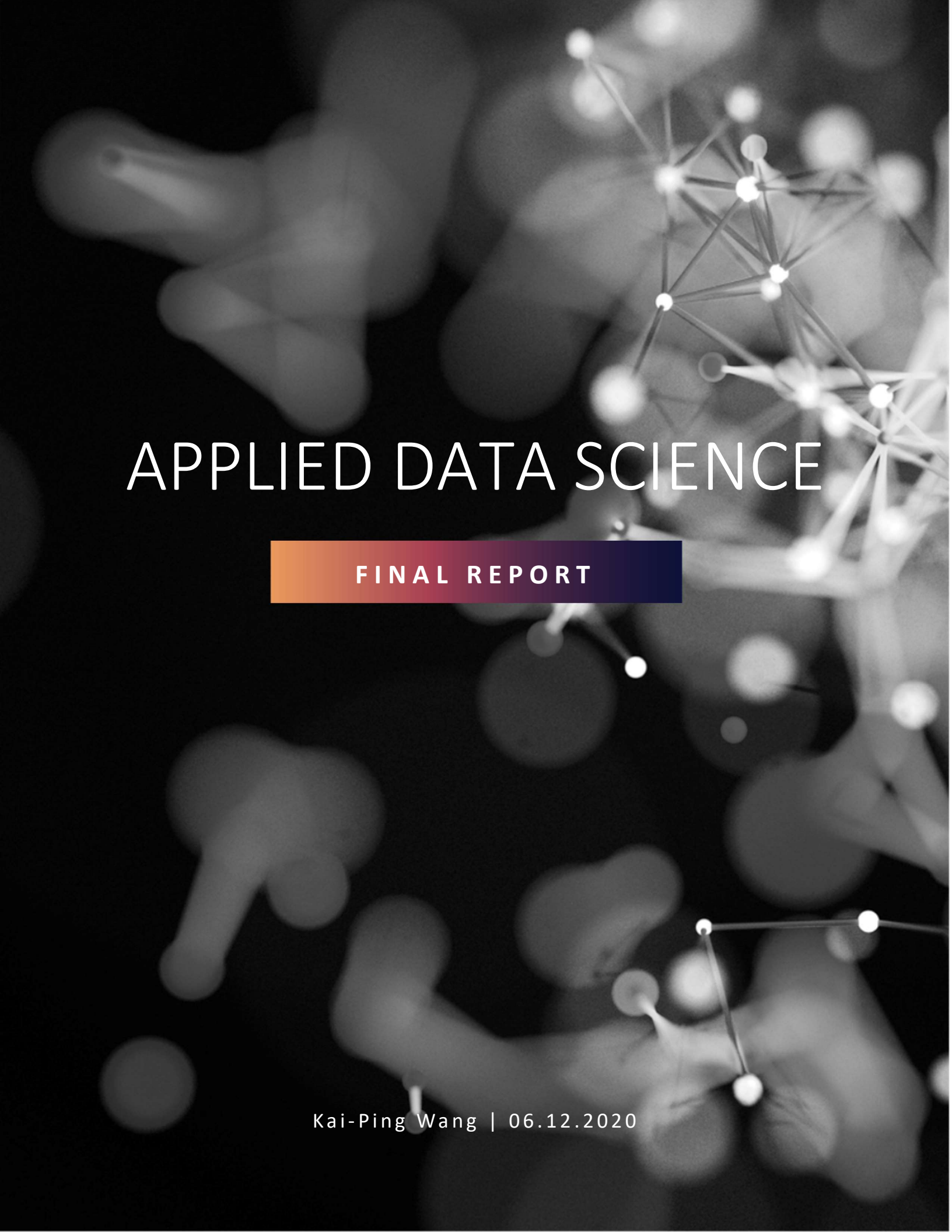
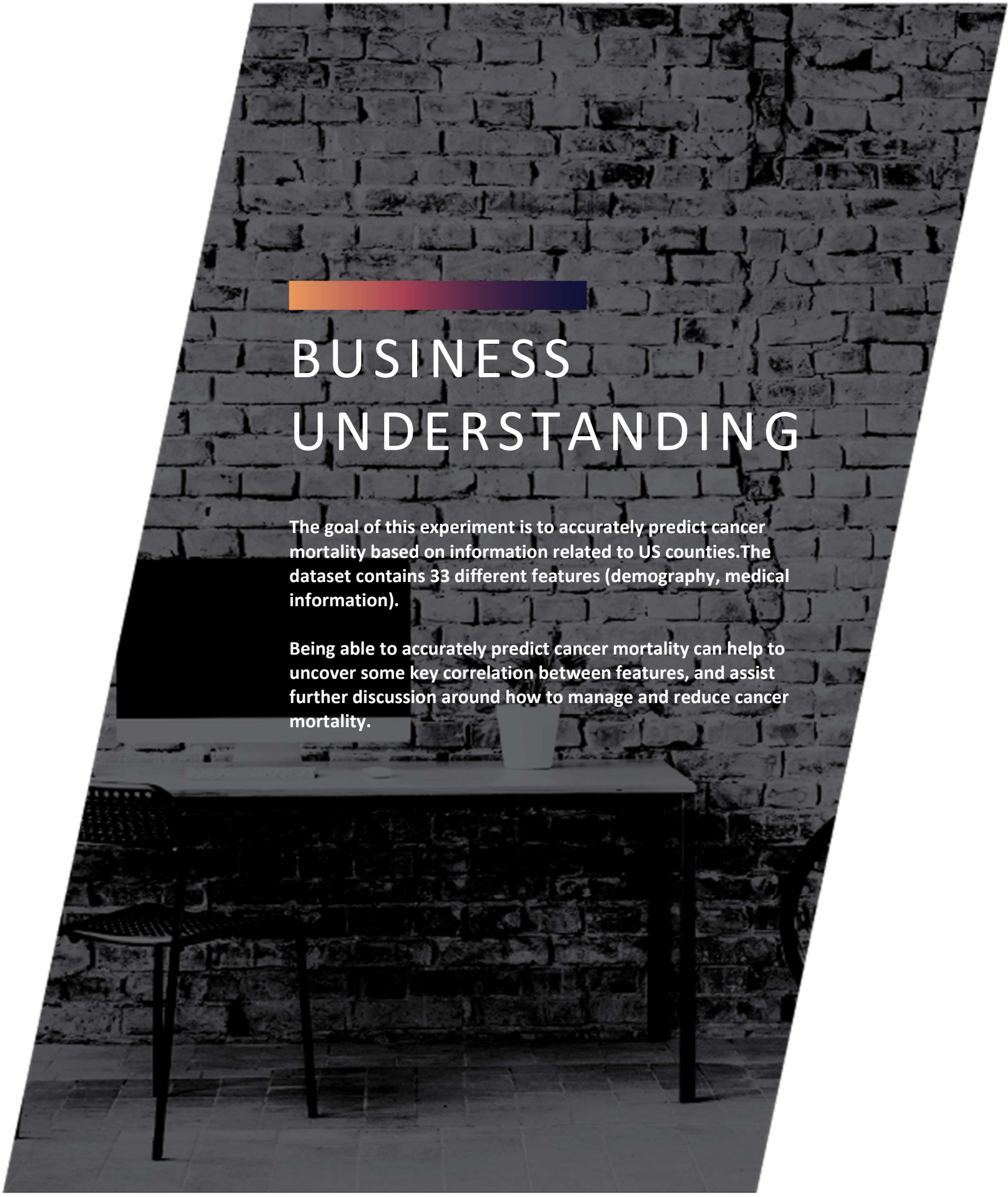


APPLIED DATA SCIENCE

FINAL REPORT

Kai-Ping Wang | 06.12.2020



BUSINESS UNDERSTANDING

The goal of this experiment is to accurately predict cancer mortality based on information related to US counties.The dataset contains 33 different features (demography, medical information).

Being able to accurately predict cancer mortality can help to uncover some key correlation between features, and assist further discussion around how to manage and reduce cancer mortality.

DATA UNDERSTANDING

DATA QUALITY

- Some features have up to 75% missing values, which makes them **NOT** suitable to be included in the prediction model.
- Some features have up to 70% potential erroneous values, these values look peculiar and almost impossible based on common sense.

  These features are still included in the modeling stage; however, may contribute to the poor performance due to the potential quality issue.

  We shall go back to the data providers to review and rectify any issues in these features.
- Some features have minor unreasonable values that seems to be typo and can be calculated back to be in line with the rest of the value.

  These features are included in the modeling stage by using data processing to bring them into reasonable value.

COLLECT DATE

The data used in this experiment is coming from https://raw.githubusercontent.com/aso-uts/applied_ds/master/assignment1/cancer_reg.csv

DESCRIBE DATA

There are 34 features and 3047 observations in this dataset. Most features are in numeric format, and Geography and binnedInc features are in string format.

#	Column	Non-Null Count	Dtype
0	avgAnnCount	3047 non-null	float64
1	avgDeathsPerYear	3047 non-null	int64
2	TARGET_deathRate	3047 non-null	float64
3	incidenceRate	3047 non-null	float64
4	medIncome	3047 non-null	int64
5	popEst2015	3047 non-null	int64
6	povertyPercent	3047 non-null	float64
7	studyPerCap	3047 non-null	float64
8	binnedInc	3047 non-null	object
9	MedianAge	3047 non-null	float64
10	MedianAgeMale	3047 non-null	float64
11	MedianAgeFemale	3047 non-null	float64
12	Geography	3047 non-null	object
13	AvgHouseholdSize	3047 non-null	float64
14	PercentMarried	3047 non-null	float64
15	PctNoHS18_24	3047 non-null	float64
16	PctHS18_24	3047 non-null	float64
17	PctSomeCol18_24	762 non-null	float64
18	PctBachDeg18_24	3047 non-null	float64
19	PctHS25_Over	3047 non-null	float64
20	PctBachDeg25_Over	3047 non-null	float64
21	PctEmployed16_Over	2895 non-null	float64
22	PctUnemployed16_Over	3047 non-null	float64
23	PctPrivateCoverage	3047 non-null	float64
24	PctPrivateCoverageAlone	2438 non-null	float64
25	PctEmpPrivCoverage	3047 non-null	float64
26	PctPublicCoverage	3047 non-null	float64
27	PctPublicCoverageAlone	3047 non-null	float64
28	PctWhite	3047 non-null	float64
29	PctBlack	3047 non-null	float64
30	PctAsian	3047 non-null	float64
31	PctOtherRace	3047 non-null	float64
32	PctMarriedHouseholds	3047 non-null	float64
33	BirthRate	3047 non-null	float64

EXPLORE DATA

When sorting the dataset based on **TARGET_deathRate** in descending order, it is noticed that the **Geography** information is also sorted in groups by state. Similar trend can be observed in features **incidenceRate** and **studyPerCap**; however, there are lots of potential erroneous values in these two features.

FORM HYPOTHESES

- **Higher the incidence rate, higher the cancer mortality.** It is easy to assume that higher the case numbers, higher the death counts.

  However, if the treatment is proper and in time, the incident rate should not be highly related to death rate. If they are highly related, then the treatment of cancer should be revisited or further investigated.

  This can have further indication on if patients receive different treatment based on what insurance they have, or if different treatments were given out based on location or race…etc

- **Based on the potential matching trend of Geography**, the health insurance coverage, medical treatment, education level and cancer awareness…etc are something to look further into.

POTENTIAL ERRONEOUS VALUE

- **studyPerCap** – has 70% observation with 0 as value. It doesn't quite make sense to have that many counties with this value. However, replacing it with other fixed value on this scale wouldn't make sense either. Leave it as is. (Will experiment with and without these value in modeling stage)
- **avgAnnCount** – has 70% observation with value 1962.667684. It doesn't make sense either. Leave it as is, and will experiment with and without in modeling stage.
- **incidenceRate** – has 70% observation with value 453.5494221. These observations are from the same observation from **avgAnnCount** as well. Will experiment with and without these value in modeling stage.

DATA PREPARATION

SELECT DATA

- Exclude **PctSomeCol18_24** as too much missing data
- Exclude **PctPrivateCoverageAlone** & **PctEmployed16_Over** as there are some missing data, and there's no apparent matching trend with **TARGET_deathRate**.

CLEAN DATA

- **AvgHouseholdSize** has some value between 0 and 1, which doesn't make sense. Hence, bring those value from float to integer by multiplying 100.
- **MedianAge** has some value larger than 100, which doesn't make sense. Hence, bring those value to number between 1 and 100 by dividing by 10.

CONSTRUCT DATA

- Split **Geography** into **State** and **County**, and do one-hot encoding on State feature to transform into dummy variables. Once it's done, exclude geography, state and county features.
- Map **binnedInc** into integers that scale from 1 to 10 based on the bracket.

FORMAT DATA

- Feature scaling using **StandardScaler** to standardize all features on training set only.
- Standardize testing set using the **StandardScaler** fitted by training set to avoid data leakage.

MODELING

HYPERPARAMETERS

- **RandomForestRegressor**
  - max_depth
  - max_leaf_nodes
  - min_samples_leaf
- **SVR**
  - Kernel
  - C
  - Epsilon

TEST DESIGN

- Split training set and test set by 80:20.
- Do cross-validation and K-fold of 5 to fully utilize the training set.

MODEL OPTIONS

Select a number of regression models as candidates, and pick the final champion in evaluation stage based on performance. Also fine tune these models further using hyperparameters.

- **Multivariate Linear Regression**
- **Random Forest Regression**
- **Support Vector Regression**

EVALUATION

Mean Square Error is used to evaluate the performance of each model included in this experiment. The goal is to find a consistent MSE across both training set and test set, and not overfitting nor underfitting. **Below is the result of all tried models, and the highlighted ones are the best of each algorithm.**

	MSE - train	MSE - test
Multivariate Linear Regression (default)	**0.4177**	**0.4075**
Random Forest Regression (default)	0.0632	0.4766
Random Forest Regression (max_depth=15)	0.0727	0.4767
Random Forest Regression (max_depth=5)	0.3833	0.5302
Random Forest Regression (max_depth=2)	0.6387	0.6871
Random Forest Regression (max_depth=5, min_samples_split=100)	0.4317	0.5454
Random Forest Regression (max_depth=5, min_samples_split=200)	0.4692	0.5562
Random Forest Regression (max_depth=5, min_samples_split=100, min_samples_leaf=100)	**0.5232**	**0.5825**
Random Forest Regression (max_depth=5, min_samples_split=100, min_samples_leaf=50)	0.4639	0.5423
Random Forest Regression (max_depth=5, min_samples_split=100, min_samples_leaf=30)	0.4495	0.5361
Support Vector Regression (default)	0.3059	0.5235
Support Vector Regression (kernel = 'linear')	**0.4155**	**0.4889**
Support Vector Regression (kernel = 'sigmoid')	20.2363	21.9596
Support Vector Regression (kernel = 'poly')	0.2771	0.6811
Support Vector Regression (kernel = 'linear', epsilon=0.5)	0.4119	0.4947
Support Vector Regression (kernel = 'linear', epsilon=1)	0.4345	0.5319
Support Vector Regression (kernel = 'linear', epsilon=0.5, C=2)	0.4121	0.4958
Support Vector Regression (kernel = 'linear', epsilon=0.5, C=0.5)	0.4118	0.4947

WHAT'S NEXT

The best model in this experiment is Multivariate Linear Regression; however, the performance is not good enough for production deployment. Although the MSE is consistent between training set and test set, it is still too high for deployment.

When removing all potential erroneous values from incidenceRate and studyPerCap, the MSE has dropped significantly and the performance is promising. This shows both features have great significance in predicting the cancer mortality accurately.

However, by doing so, the meaningful dataset has dropped to 30% only. If there are some issues in capturing these features in any counties or in the system, then the future prediction is still not reliable as values of these two features may be erroneous.

The recommended next step is to discuss with data provider to review both features (incidenceRate and studyPerCap) and rectify any issues if there are. Once this is completed, rerun the model with latest data and check the performance again.