# LionBase x MedX

Data Report - May 18th, 2020

# Summary of Results

**To distribute the new drug to the most number of patients possible, MedX should prioritize distribution to counties that have a high incidence rate, lower median income, and higher rate of unemployment (i.e. less affluent communities).**

- These communities tend to have a much higher cancer mortality per capita compared to wealthier communities — they have a much greater need for this new drug
- Targeting these areas will result in the greatest number of lives being saved; the drug will be going to the areas with greatest demand
- Age, race, and marital status are nearly useless for predicting which counties will have the highest demand for the drug

# Data Output

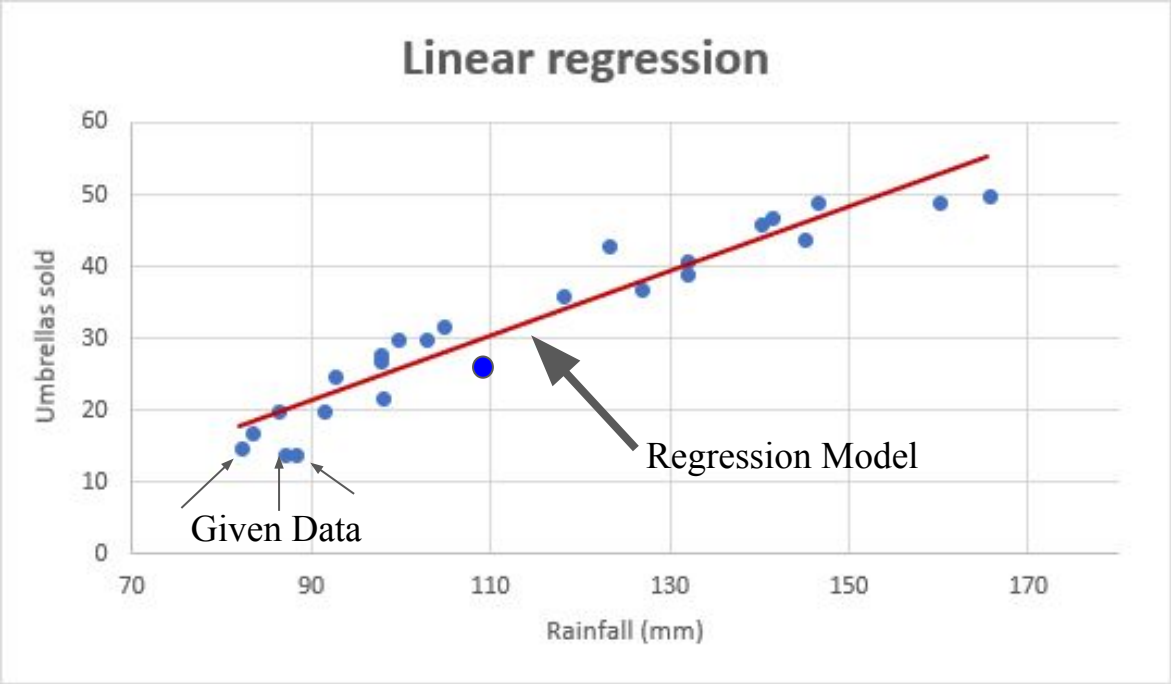
In the project zip file, you will find the following:

- Data (folder)
  - cancer_data.csv
- Regression Model (folder)
  - LionBaseXMedX.ipynb

# Methods

1. Data used in this model came directly from the “cancer_data.csv” file provided to LionBase
2. Data was cleaned prior to analysis:
    a. Null values were identified and addressed
    b. Outliers were taken out
    c. Duplicate entries were checked for
    d. Redundant columns were taken out
    e. Separate standardized data set was created
3. An initial linear regression model was fitted to the data
4. Initial model was iterated over to create an improved linear regression model

# Linear Regression Example:



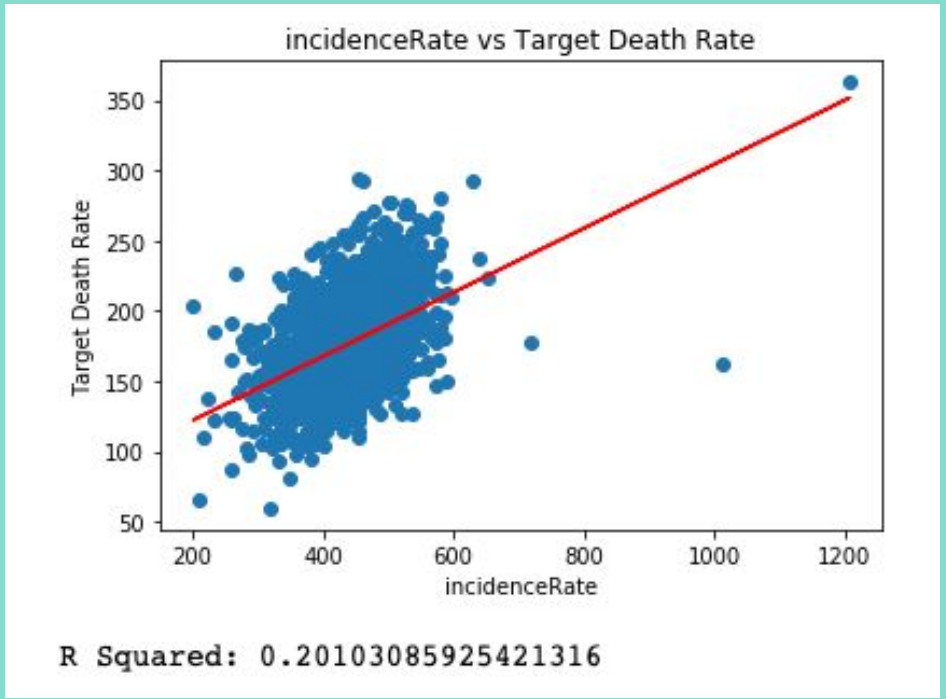

- Create a model based on given data
- Optimize the model to minimize the error in predictions
- Can be used to predict outcomes for future events

# Analysis/Visualization

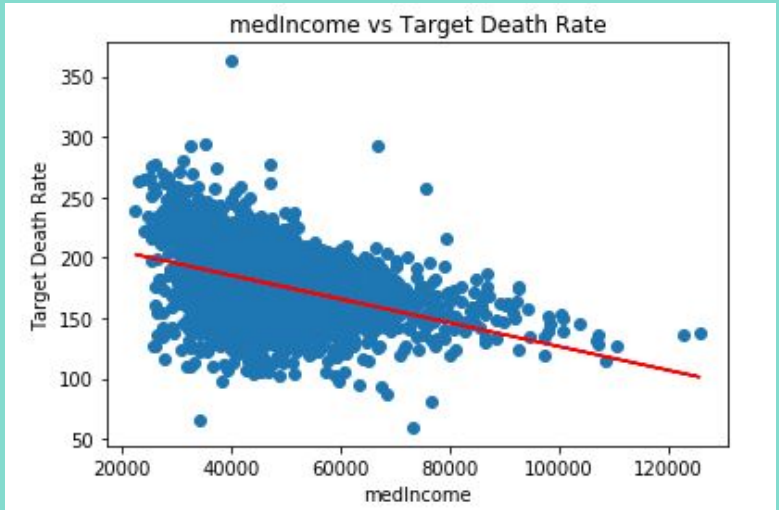
- Created regression model is multidimensional
- Represented with a collection of 2D scatterplots (one per predicting variable)

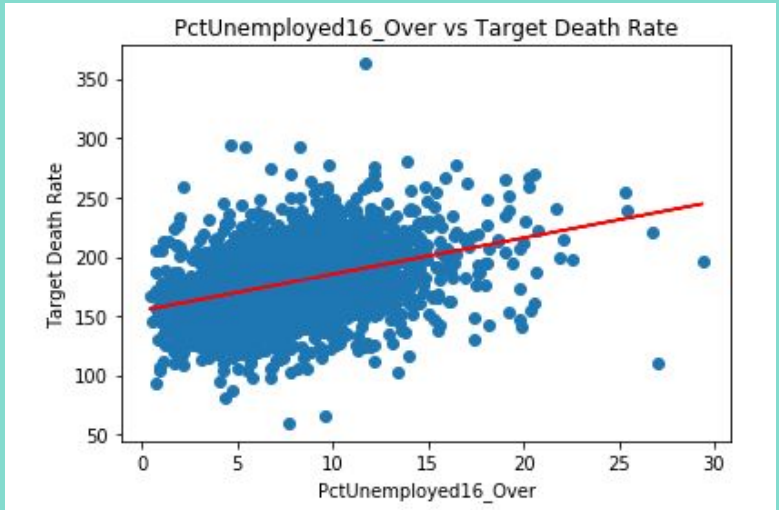
**Conclusion:** Counties with higher amounts of cancer diagnoses per capita tend to have higher cancer mortality rates

**Action:** Prioritize shipping the drug to areas with high incidence rates to maximize the number of lives saved



# Further Analysis of Model:

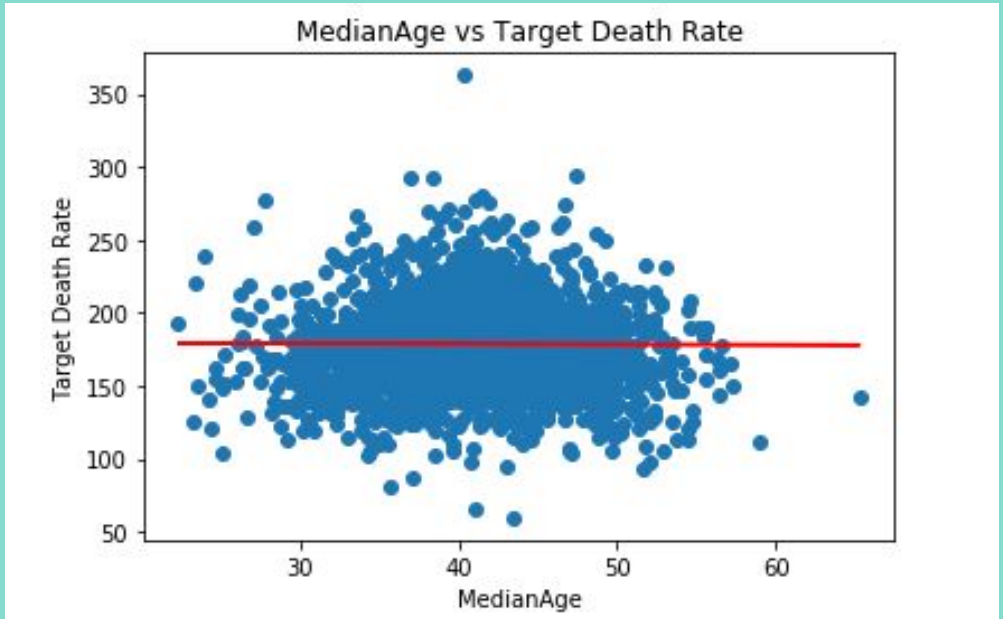



**Conclusion:** Counties with lower median incomes and higher rates of unemployment have greater death rates due to cancer

**Action:** Focus on shipping treatments to these areas where demand is greatest

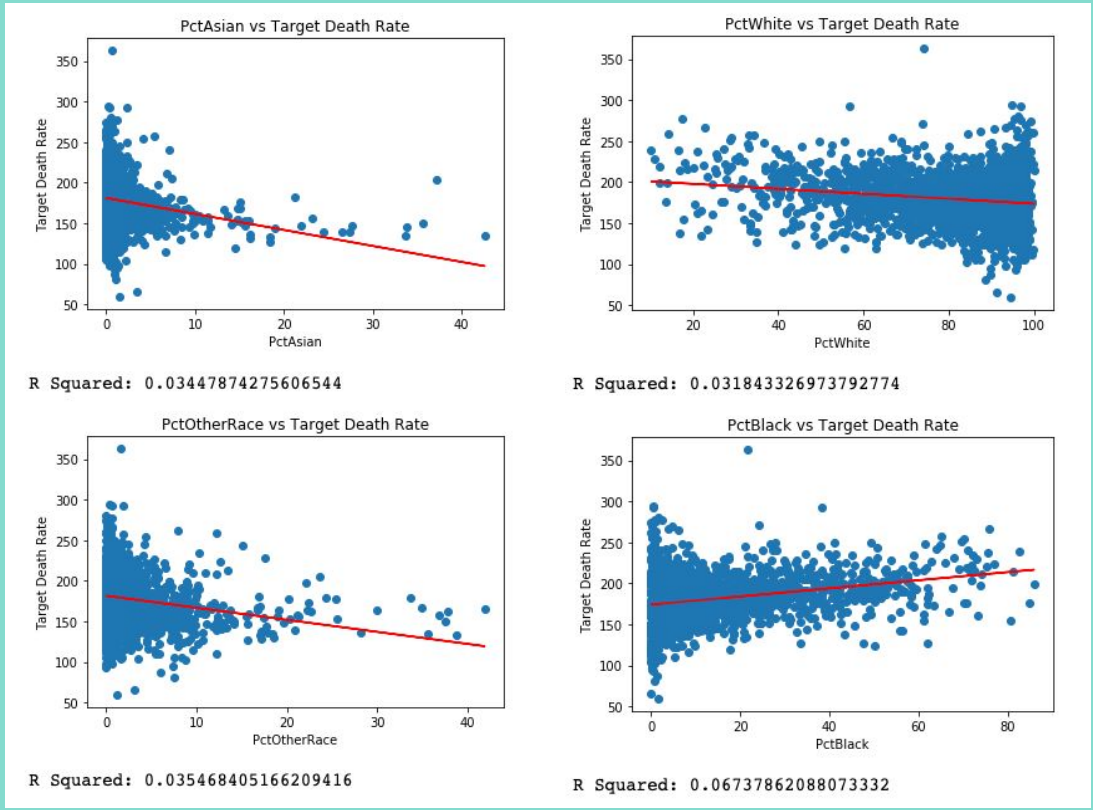
# Other Findings:




- The age of a county's population does not make a difference in the death rate due to cancer
- Model line is almost perfectly flat

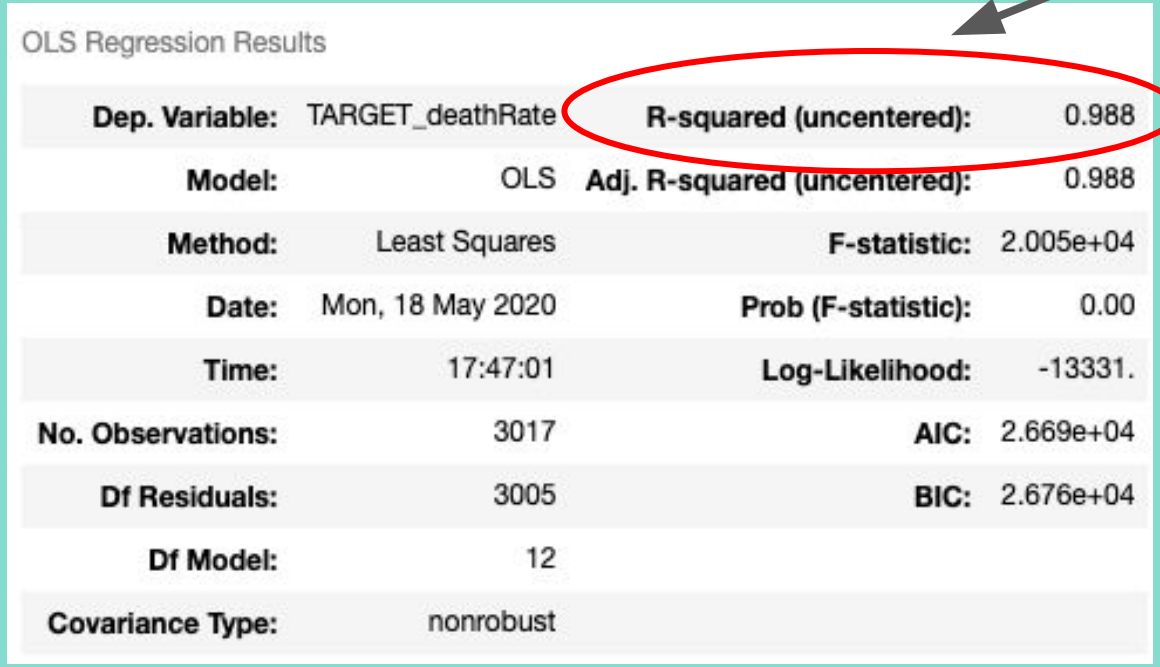
# Other Findings:

- Demographics of counties are not strong predictors
- These statistics are not relevant when determining optimal distribution of the drug



# Model Quality

98.8% of variance in data is explained by this model



OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | TARGET_deathRate | **R-squared (uncentered):** | 0.988 |
| **Model:** | OLS | **Adj. R-squared (uncentered):** | 0.988 |
| **Method:** | Least Squares | **F-statistic:** | 2.005e+04 |
| **Date:** | Mon, 18 May 2020 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 17:47:01 | **Log-Likelihood:** | -13331. |
| **No. Observations:** | 3017 | **AIC:** | 2.669e+04 |
| **Df Residuals:** | 3005 | **BIC:** | 2.676e+04 |
| **Df Model:** | 12 | | |
| **Covariance Type:** | nonrobust | | |

# Key Takeaways

1. Prioritize shipping to counties that have high incidence rates, lower median income, and higher rate of unemployment
2. Using the linear regression model, we can quickly determine which counties are at higher risk and can thus easily determine the optimal distribution
3. Determining optimal shipping orders can be done without incurring any additional costs
4. Ensures that as many lives are saved as possible and maximum revenue is achieved

# Next Steps

- Pass all U.S. counties through the linear regression model to create a comprehensive priority listing of all potential counties to ship to
- Create algorithm to determine the optimal path to various counties, maximizing the quantity of products delivered while minimizing shipping costs