CSE 4/574 - Introduction to Machine Learning

Assignment 2, Summer 2025

Jayant Som
MS Student in Quantum Science & Nanotechnology
Department of Electrical Engineering
University at Buffalo
Buffalo, NY 14068
jsom@buffalo.edu

Abstract

This report offers a concise introduction to key machine learning concepts, emphasizing data visualization, linear regression, and regularization methods - specifically Ridge Regression and Gradient Descent. It begins by illustrating how visual exploration of the data aids in understanding dataset characteristics, identifying trends, and preparing for model training. The study applies Linear Regression, Ridge Regression, and Gradient Descent to predict insurance charges using a publicly available dataset, starting with exploratory data analysis to examine relationships between variables such as age, BMI, and smoking status and their influence on insurance costs. Three regression models are implemented from scratch (without scikit-learn): Ordinary Least Squares, solved analytically through a closed-form solution; Ridge Regression, which introduces L2 regularization and involves hyperparameter tuning for optimal penalty; and Gradient Descent, an iterative optimization technique used to minimize the loss function.

1 PART I: DATA ANALYSIS

1.1 NETFLIX Dataset

This dataset provides details about Netflix's content library, including both movies and TV shows, along with metadata such as:

- Title, director, and cast.
- Country of origin, release year, and date added to Netflix.
- Rating, duration, and genre/category.
- Description (a brief summary of the content)

The dataset contains structured data in tabular form with rows (entries) and columns (attributes).

1.1.1 Dataset Details and Statistics:

a. Shape:

- Total rows: 8807
- Total Columns: 12 (show_id, type, title, director, cast, country, date_added, release_year, rating, duration, listed_in, description)

b. Data-Types:

- This dataset contains mixed data-types:
  - Categorical: type (Movie/TV Show), rating, country, listed_in

(genre)
- Textual: title, director, cast, duration (in minutes or seasons), description
- Date/Time: date_added, release_year
- Identifier: show_id (unique key for each entry)

c. Statistics:

- The dataset contains only one column with numerical values: release_year

  Mean release year is 2014.
  Standard deviation is approx. 9 years.
  Earliest release year is 1925 and latest release year is 2021.

- For columns containing string data, below are the meaningful top (most frequent) statistics:

  Top content type: Movies
  Top cast: David Attenborough
  Top country: United States
  Top rating: TV-MA
  Top duration: 1 season
  Top category: Dramas, International Movies

d. Missing values:

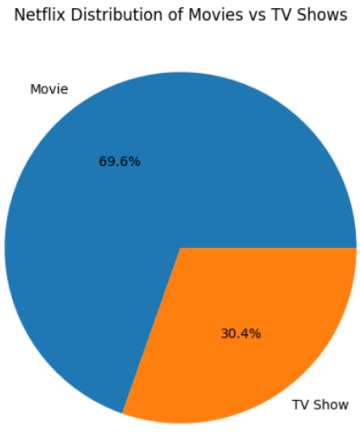
The dataset contains missing values in multiple columns:

Director: 2643 entries
Cast: 915 entries
Country: 831 entries
Date added: 10 entries
Rating: 4 entries
Duration: 3 entries

1.1.2 Visualizations:

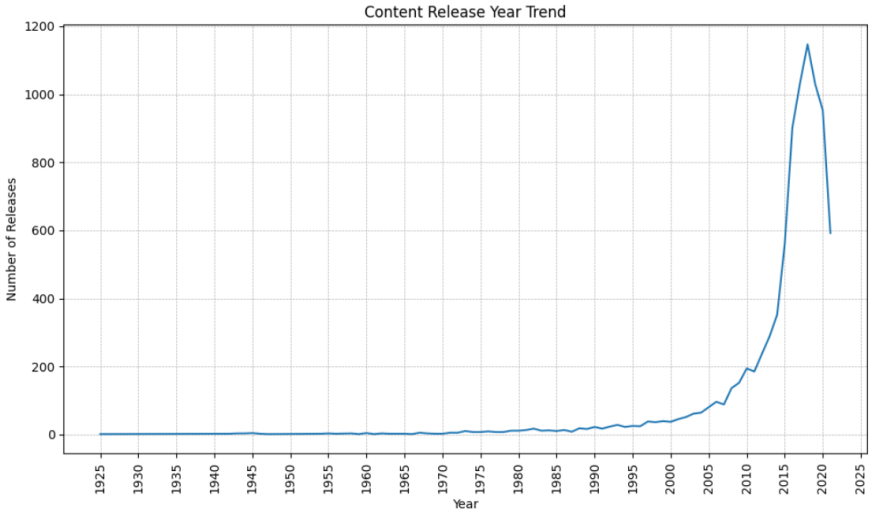
Libraries used: Matplotlib, Seaborn, Squarify, Wordcloud

a. Distribution of Movies vs TV Shows



Observations: The distribution of content types in the dataset reveals that movies constitute 69.6% of the total entries, while TV shows account for the remaining 30.4%. This indicates a significant predominance of movies over TV shows in the Netflix catalog represented by this dataset.

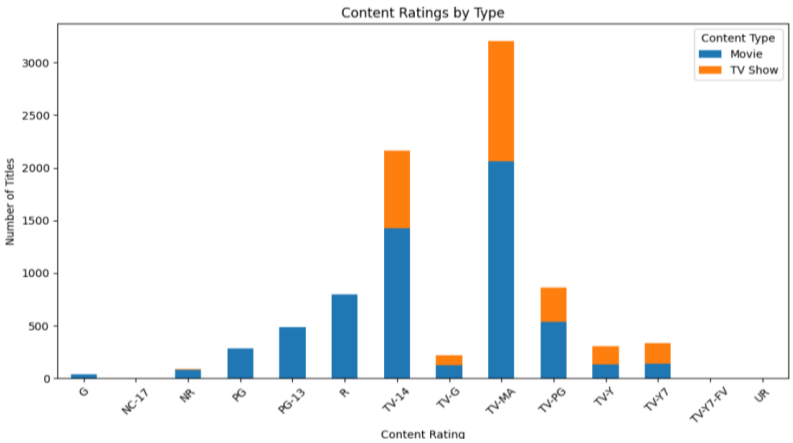
b. Release Year Trend



Observations:

- Pre-1990s: Minimal content releases, likely due to limited digital media adoption and archival constraints.
- 1990-2005: Gradual increase in releases, coinciding with early digital media expansion and improved content production.
- 2005 Onwards: Sharp upward trend, correlating with widespread internet adoption and the rise of streaming platforms.
- Peak (2018-2019): Highest release volumes observed just prior to the COVID-19 pandemic, reflecting industry growth.
- COVID-19 Impact (2019-2020): Significant decline in new releases due to production delays and disruptions.
- Post-2020: Continued downward trend, possibly indicating long-term pandemic effects or market saturation.

c. Content Rating Distribution by Type



Ratings assigned by the MPAA (Motion Picture Association of America) for films and the TV Parental Guidelines Monitoring Board for television content:

- G - General Audiences
- NC-17 - No Children Under 17 Admitted
- NR - Not Rated
- PG - Parental Guidance Suggested
- PG-13 - Parents Strongly Cautioned (Under 13 Requires Parental Guidance)
- R - Restricted (Under 17 Requires Accompanying Parent or Adult Guardian)

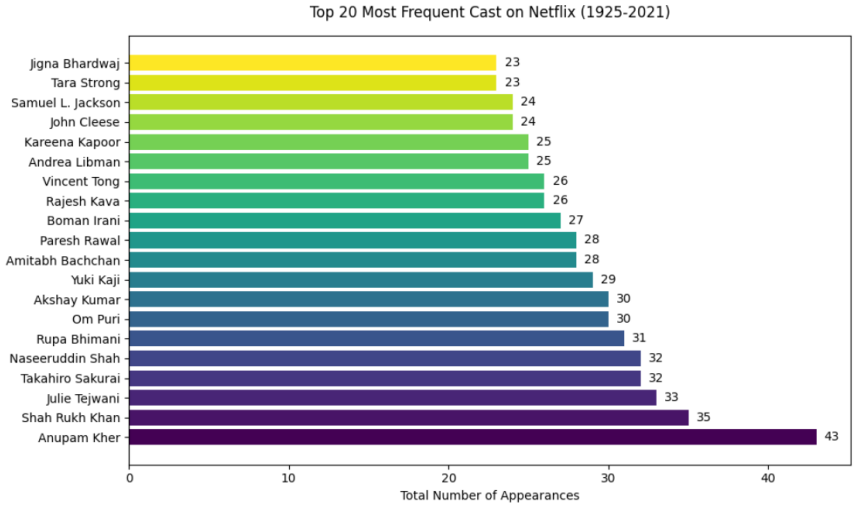
- TV-14 - Parents Strongly Cautioned (Unsuitable for Children Under 14)
- TV-G - General Audience (Suitable for All Ages)
- TV-MA - Mature Audience Only (Designed for Adults)
- TV-PG - Parental Guidance Recommended
- TV-Y - Designed for Young Children (All Children)
- TV-Y7 - Directed to Older Children (Ages 7 and above)
- TV-Y7-FV - Directed to Older Children (Fantasy Violence)
- UR - Unrated

Observations: Most prevalent ratings across both movies & TV shows are:

- TV-MA (Mature Audience Only) - Indicating content intended for adults.
- TV-14 (Parents Strongly Cautioned - Unsuitable for Children Under 14) - Suggesting teen and young adult-oriented material.

This trend highlights that adult-oriented content dominates Netflix's catalog, with a significant portion of titles targeting mature audiences rather than younger or family-friendly demographics.

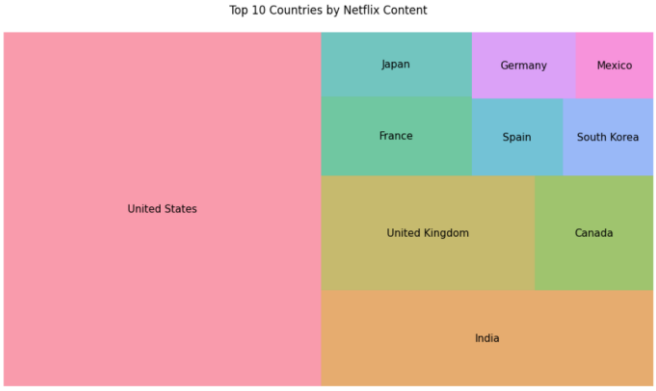
d. Most Frequent Cast



Observations: Top Cast are:

- Anupam Kher (Actor from India)
- Shah Rukh Khan (Actor from India)
- Julie Tejwani (Voice Actor from India)
- Takahiro Sakurai (Voice actor from Japan)
- Naseeruddin Shah (Actor from India)

e. Top Countries:

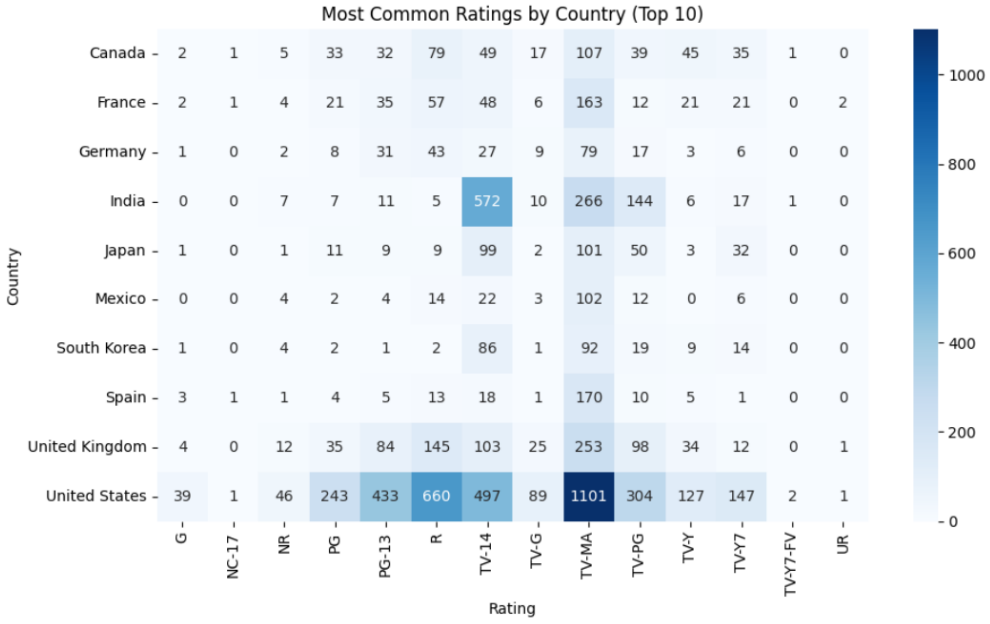


Observations: The countries with most Netflix content are USA, India & UK

f. Most Common Genres:

Observations: The content category most available in dataset are TV Shows, International Movies and International Drama.

g. Analysis of Ratings by Country



Observations: Netflix's library is adult-centric (TV-MA/R) in Western markets, while India and Japan show unique deviations - India with teen-focused (TV-14) and Japan with dual adult/kids (R/TV-Y) emphasis.

1.2 INSURNCE Dataset

This dataset includes variables that help predict insurance charges based on factors such as:

- Demographics: Age, sex
- Health Metrics: Body Mass Index (BMI)
- Lifestyle Choices: Smoking status
- Family Information: Number of children
- Geographical Region: Region of residence in the U.S.

The dataset contains structured data in tabular form.

1.2.1 Dataset Details and Statistics:

a. Shape:

- Total rows: 1338,
- Total Columns: 7 (age, sex, bmi, children, smoker, region, charges)

b. Data-Types:

This dataset contains mixed data-types:

- Numerical (Continuous): age, bmi, charges
- Numerical (Discrete): children
- Categorical/Nominal: sex (male/female), smoker (yes/no), region (geographic region: northeast, northwest, southeast, southwest)

c. Statistics:

- The dataset contains only multiple columns with numerical values:
  Age: Mean age is 39 yr. with standard deviation of 14 yr.
  BMI: Mean bmi is 30.7 with standard deviation of 6.1.
  Charges: Mean charge is $13270 with standard deviation of $12110.

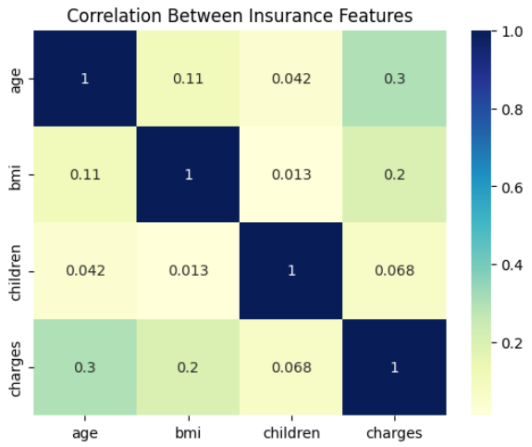
- For columns containing string data, below are the meaningful top (most frequent) statistics:
  Top sex: Male
  Top smoking-status: No
  Top region: Southeast

d. Missing values:

This dataset does not contain any missing values.

1.2.2 Visualizations:

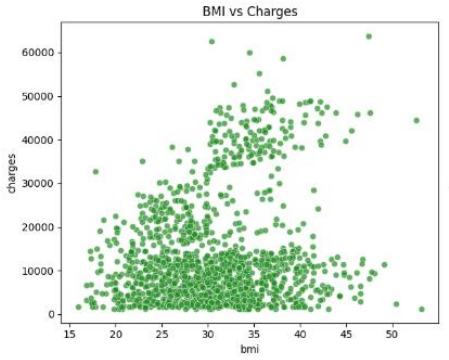
a. Correlation Matrix of Numerical Features

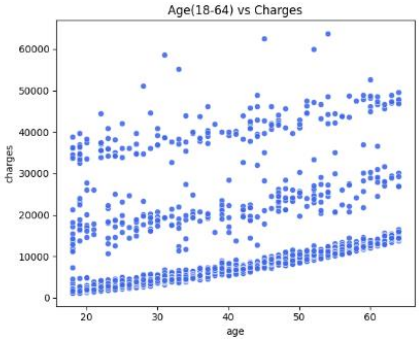


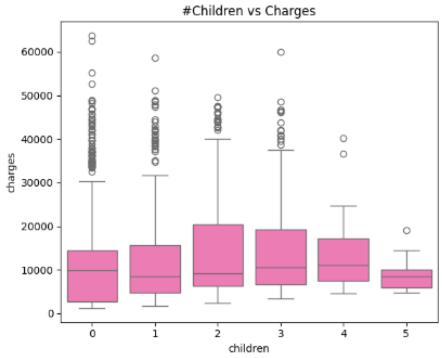
Observations:

- The strongest correlation is between age and charges (0.30)
- BMI has weaker but still notable correlation with charges (0.20)
- # children shows almost no linear relationship with charges
- No strong multicollinearity between predictors (all < 0.3)

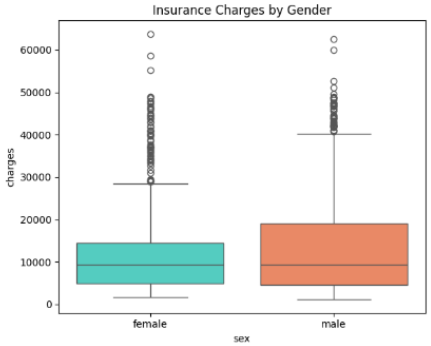
b. Numerical Features vs Target (Charges)

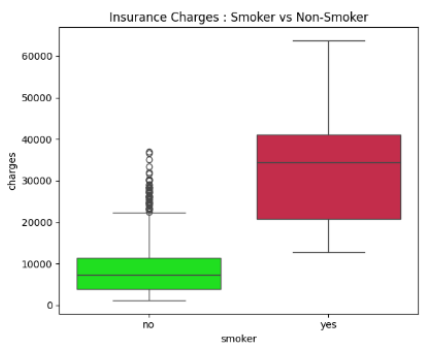


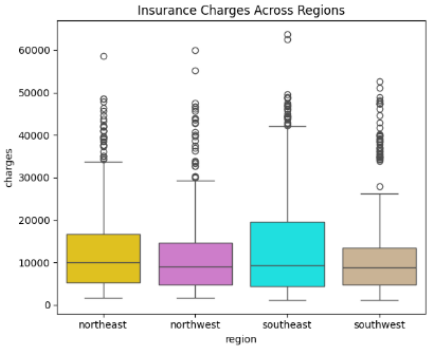




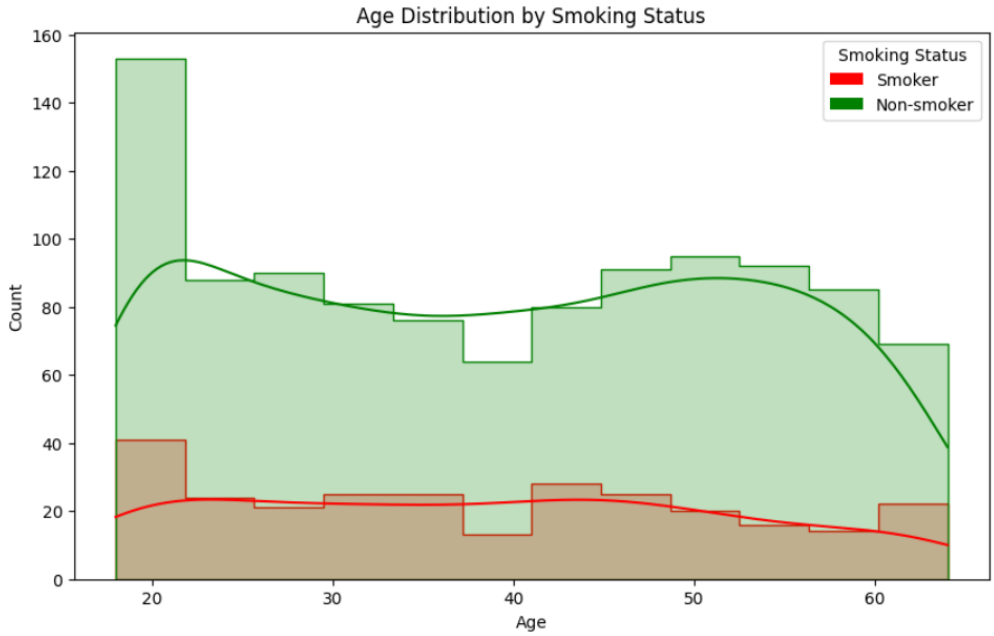
c. Categorical Features vs Target (Charges)





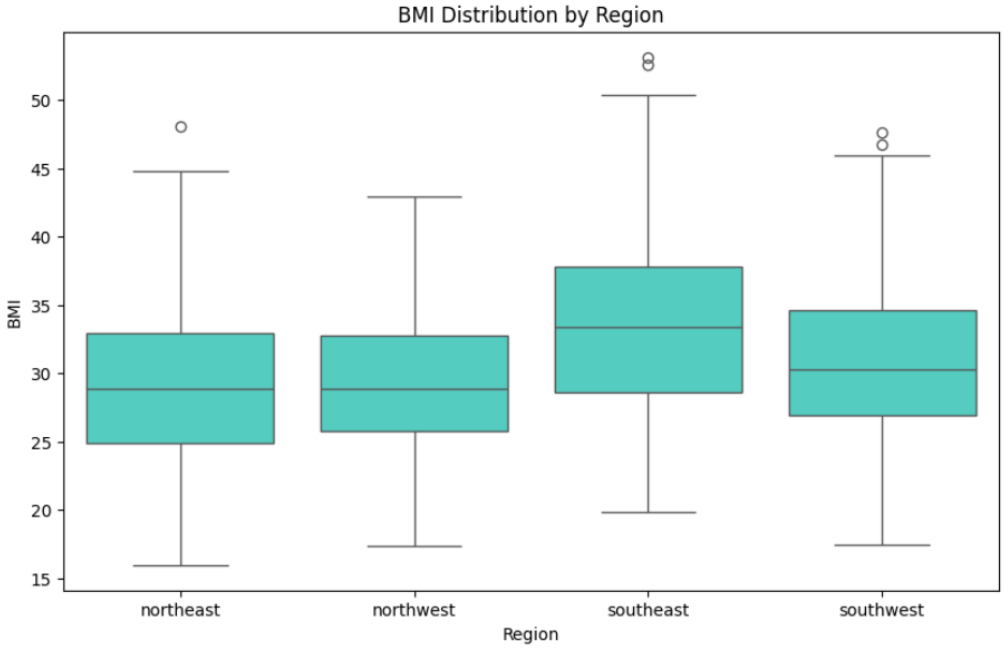


d. Age Distribution by Smoker Status



Observations: Smokers tend to be slightly younger than non-smokers on average. Most smokers are in their 20s-50s, while non-smokers are more evenly spread across all ages.

e. BMI Distribution by Region



Observations:

- The Southeast stands out as having the heaviest residents
- Southwest residents seem healthiest weight-wise
- Northern regions are pretty similar to each other

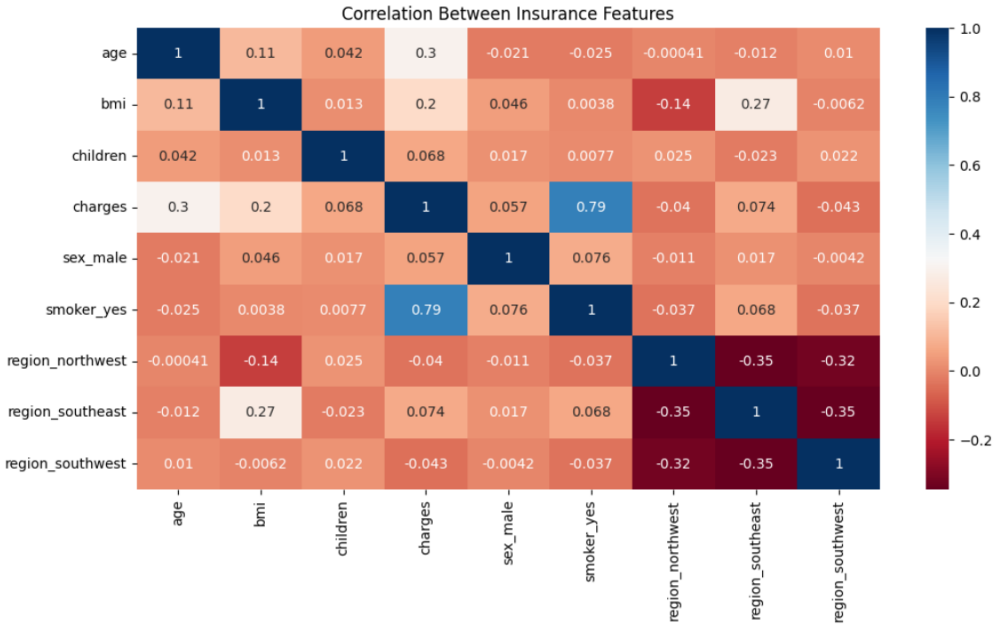
This helps explain why insurance costs vary by region: heavier people often have more health issues and require more insurance.

2 PART II: LINEAR REGRESSION

2.1 Feature Engineering

- Converted the string features (sex, smoker, region) to categorical using get_dummies() function of Pandas.
- Normalized non-categorical features (age, bmi, children) using max-min normalization technique.
- Converted whole dataset into float data-type.

2.2 Correlation



Observations:

- The strongest correlation is between smoker_yes and charges (0.79)
- Age has notable correlation with charges (0.3)
- BMI has weaker but still notable correlation with charges (0.20)

2.3 Feature and Target Separation (N=1338)

- Target y was chosen to be charges. Shape of y: (1338,)
- Features X were all other columns other than charges. Total 8 features, so d = 8. Shape of X: (1338, 8)

2.4 Adding Implicit bias

- Added a column of 1s for the intercept i.e., bias wo. Now X has shape (N, d+1) i.e., new shape of X: (1338, 9)

2.5 Train-Test Split

- Divided the dataset into training and test, as 80% - training, 20% - testing.
- X_train shape: (1070, 9)
  y_train shape: (1070,)
  X_test shape: (268, 9)
  y_test shape: (268,)

2.6 Linear Regression:

2.6.1 Calculation of weights using OLS formula

- Calculated the weights using OLS equation:

$$\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

- Linear Regression Weights vector:

  Intercept (w0): -1779.4677
  age (w1): 11936.5531
  bmi (w2): 12607.6535
  children (w3): 2195.8382
  sex_male (w4): -253.8023
  smoker_yes (w5): 23653.4332
  region_northwest (w6): -471.6674
  region_southeast (w7): -1119.2883
  region_southwest (w8): -1245.9259

2.6.2 Predictions

- Made predictions on test set

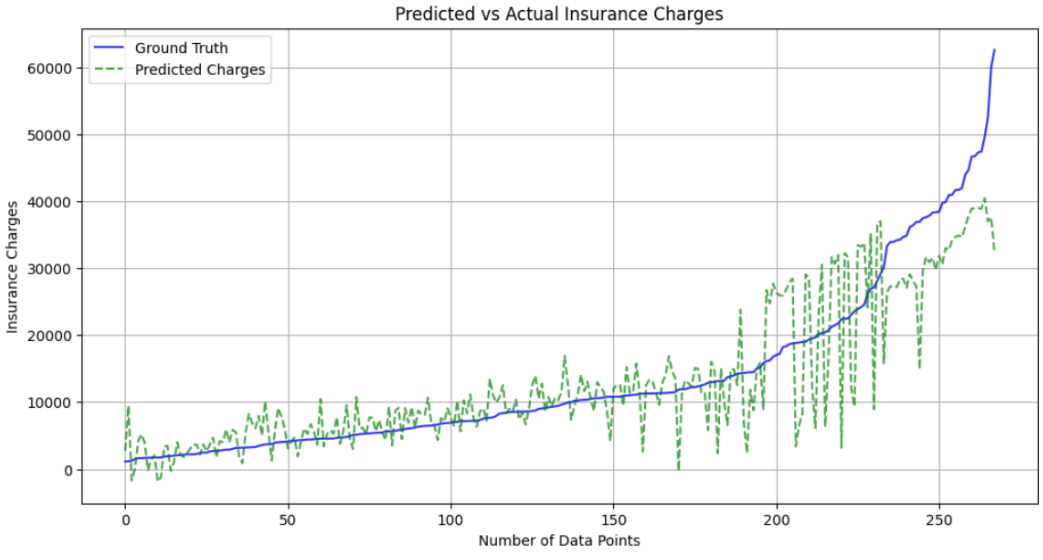
2.6.3 RMSE Calculations

- Calculated the value of RMSE using the formula:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \boldsymbol{w}^T x_i)^2}$$

- **Linear Regression RMSE: 6097.21**

2.6.4 Plot: Actual Vs Predicted



2.6.5 OLS: Benefits and Limitations

Benefits:

- It gives closed-form solution. The solution it provides is globally optimal.
- Because of squared-error loss function, there is no cancelling-out of positive and negative errors and there is heavy penalty for large errors.
- OLS computes weights analytically avoiding iterative optimization.
- Coefficients directly indicate feature importance (unit change in X affects y by w).
- OLS does not require tuning regularization strength.

Limitations:

- It is sensitive to multicollinearity – If XTX is singular (highly correlated features), it becomes non-invertible, which makes the weights unstable.
- There can be overfitting – it is prone to high variance with many features as there is no regularization present.
- It is computationally expensive for large X.
- It is assumption dependent; it requires linear relation between X and y and no autocorrelations.
- We cannot perform feature selection.

3 PART III: RIDGE REGRESSION

3.1 Calculation of weights using OLS equation for ridge-regression

- Calculated the weights using OLS equation:

$$w = (X^T X + \lambda I)^{-1} X^T y$$

I also ensured that the intercept term (bias weight w_0) is not penalized.

3.2 Finding the best penalty(lambda) value:

- Iterated for the following list of lambda values [0.001, 0.01, 0.1, 1, 10, 100, 1000, 5000, 1000000] to find the least RMSE between y_test and predictions.
- Got the following values:
  Lambda: 0.001, RMSE: 6097.21
  Lambda: 0.01, RMSE: 6097.23
  Lambda: 0.1, RMSE: 6097.47
  Lambda: 1.0, RMSE: 6100.50
  Lambda: 10.0, RMSE: 6176.85
  Lambda: 100.0, RMSE: 7591.30
  Lambda: 1000.0, RMSE: 11171.18
  Lambda: 5000.0, RMSE: 12124.17
  Lambda: 1000000.0, RMSE: 12408.47

- **Best Ridge Regression penalty: 0.001**
- **Best Ridge Regression RMSE: 6097.21**

  As lambda is increasing, the performance is decreasing. This means regularization is not possible here.

3.3 Weight calculation using best lambda value:

- Ridge Regression Weights vector:

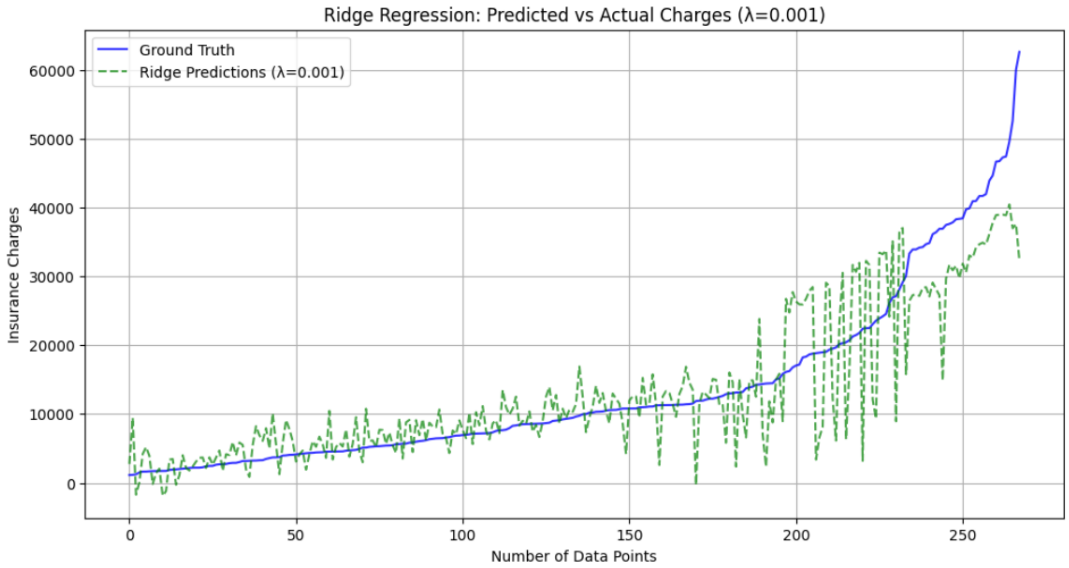
  Intercept (w0): -1779.2377
  age (w1): 11936.4648
  bmi (w2): 12607.1815
  children (w3): 2195.8154
  sex_male (w4): -253.7905
  smoker_yes (w5): 23653.2884
  region_northwest (w6): -471.6616
  region_southeast (w7): -1119.2175
  region_southwest (w8): -1245.8982

3.4 Analysis

- Ridge regression can only decrease the slope of the hyperplane but cannot increase it.
- Increasing λ is hurting performance and it is becoming highly underfitting. This means no significant overfitting was present in the linear regression. This tells us that the linear regression was already generalizing well.

- Ridge approaches OLS as λ -> 0. For λ=0.001, the regularization effect is negligible, so weights remain same as OLS estimates.
- Ridge excels when correlated and useless features exist. In this dataset, there was no correlated or useless features, so ridge's shrinkage will provide no benefit.

3.5 Plot: Actual Vs Predicted



3.6 Difference between linear and ridge regression

Aspect	Linear Regression	Ridge Regression
Objective	Minimize sum of squared errors.	Minimize sum of squared errors + penalty × (sum of squared weights).
Regularization	None, so can overfit	L2 penalty (shrinks weights toward zero).
Multicollinearity	Fails if XTX is singular	Stabilizes weights by adding λ to diagonal.
Bias-Variance Tradeoff	Low bias, high variance	Higher bias, lower variance

3.7 Main motivation of using ridge regression

- Prevents Overfitting by shrinking large weights.
- Penalizes corelated and useless features.
- Handles Multicollinearity by stabilizing solutions when features are correlated by making XTX invertible.

4 PART IV: GRADIENT DESCENT

4.1 Calculation of updated weights using Gradient Descent Method

- Calculated the weights using the update rule:

$$\boldsymbol{w} = \boldsymbol{w} - \alpha \nabla J(\boldsymbol{w})$$

where,

- alpha is the learning rate
- the Gradient equation is:

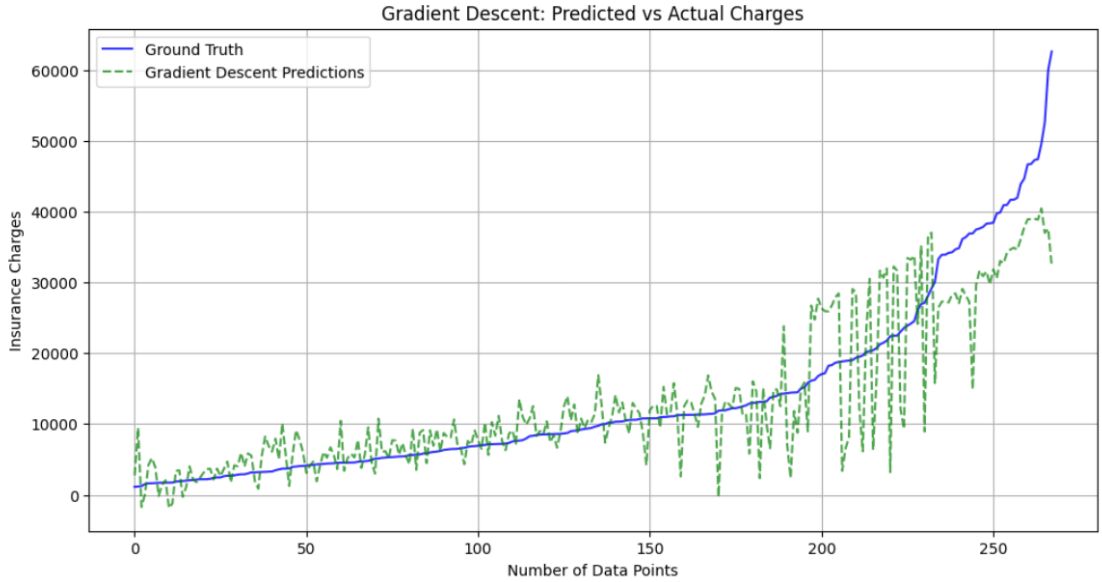
$$\nabla J(\boldsymbol{w}) = -2\boldsymbol{X}^\top \boldsymbol{y} + 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w}$$

In the above gradient equation, I also added the ridge penalty term.

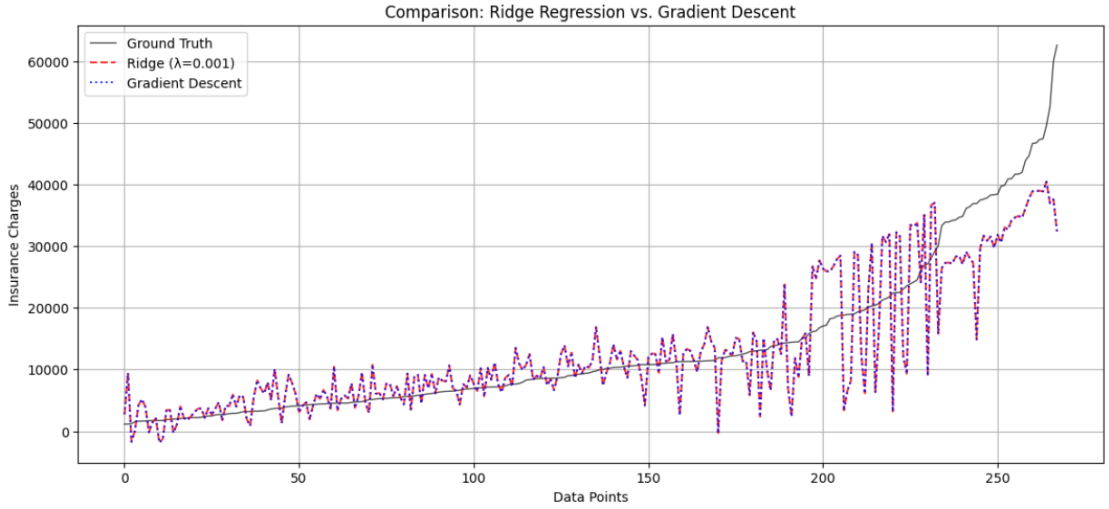
4.2 Finding the learning rate and number of iterations to get best RMSE:

- Initialized the weight vector (w) to zeros as a common starting point for optimization.
- Iterated for the following:
  - alpha values [0.0001, 0.001, 0.01, 0.1]
  - max_iter [100, 1000, 10000]
- Got the following optimal values:
  - Alpha = 0.0001
  - max_iter = 10000. Converged after 3928 iterations.
- Gradient Descent RMSE: 6097.21
- Gradient Descent Weights vector:
  Intercept (w0): -1779.2160
  age (w1): 11936.4578
  bmi (w2): 12607.1589
  children (w3): 2195.8098
  sex_male (w4): -253.7938
  smoker_yes (w5): 23653.2864
  region_northwest (w6): -471.6687
  region_southeast (w7): -1119.2220
  region_southwest (w8): -1245.9042

4.3 Plot: Actual Vs Predicted



4.4 Comparison of Ridge Vs Gradient Descent



4.5 Analysis

- Gradient Descent (with proper tuning) minimizes the same loss function as OLS. At convergence, GD's weights should match OLS weights, leading to identical RMSE. That is the global optimum.
- Small LR (0.0001) was conservative, ensuring stable convergence and no overshooting but required many iterations (3928) to converge.
- As L2 regularization had no effect, so the Gradient Descent also has no effect on the RMSE value. This tells us that the linear regression was already generalizing well.

Thus, no improvement was possible because OLS is already the optimal linear unbiased estimator for the given dataset.

5 ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Prof. Nitin Kulkarni and TA. Xiaofeng Chen (CSE Department, University at Buffalo) for their invaluable guidance, insightful feedback, and continuous support throughout this project. Their expertise and encouragement were instrumental in helping me understand the theoretical and practical aspects of linear regression, ridge regression, and gradient descent.

I also extend our thanks to the University at Buffalo for providing the resources and infrastructure necessary to complete this work.

Jayant Som

6 REFERENCES USED

a. https://pandas.pydata.org/docs/user_guide/index.html
b. https://machinelearningknowledge.ai/matplotlib-heatmap-complete-tutorial-for-beginners/
c. https://seaborn.pydata.org/tutorial.html
d. https://seaborn.pydata.org/generated/seaborn.color_palette.html
e. https://matplotlib.org/stable/gallery/subplots_axes_and_figures/index.html
f. https://seaborn.pydata.org/tutorial/regression.html
g. https://python-graph-gallery.com/590-advanced-treemap/
h. https://www.datacamp.com/tutorial/wordcloud-python
i. https://en.wikipedia.org/wiki/Motion_Picture_Association_film_rating_system
j. https://www.statology.org/ols-regression-python/
k. https://www.datacamp.com/tutorial/simple-linear-regression
l. https://www.geeksforgeeks.org/what-is-regression-analysis/#
m. https://discovery.cs.illinois.edu/guides/Statistics-with-Python/rmse/
n. https://www.geeksforgeeks.org/python/solving-linear-regression-without-using-sklearn-and-tensorflow/
o. https://www.geeksforgeeks.org/machine-learning/implementation-of-ridge-regression-from-scratch-using-python/
p. https://towardsdatascience.com/implementing-gradient-descent-in-python-from-scratch-760a8556c31f/