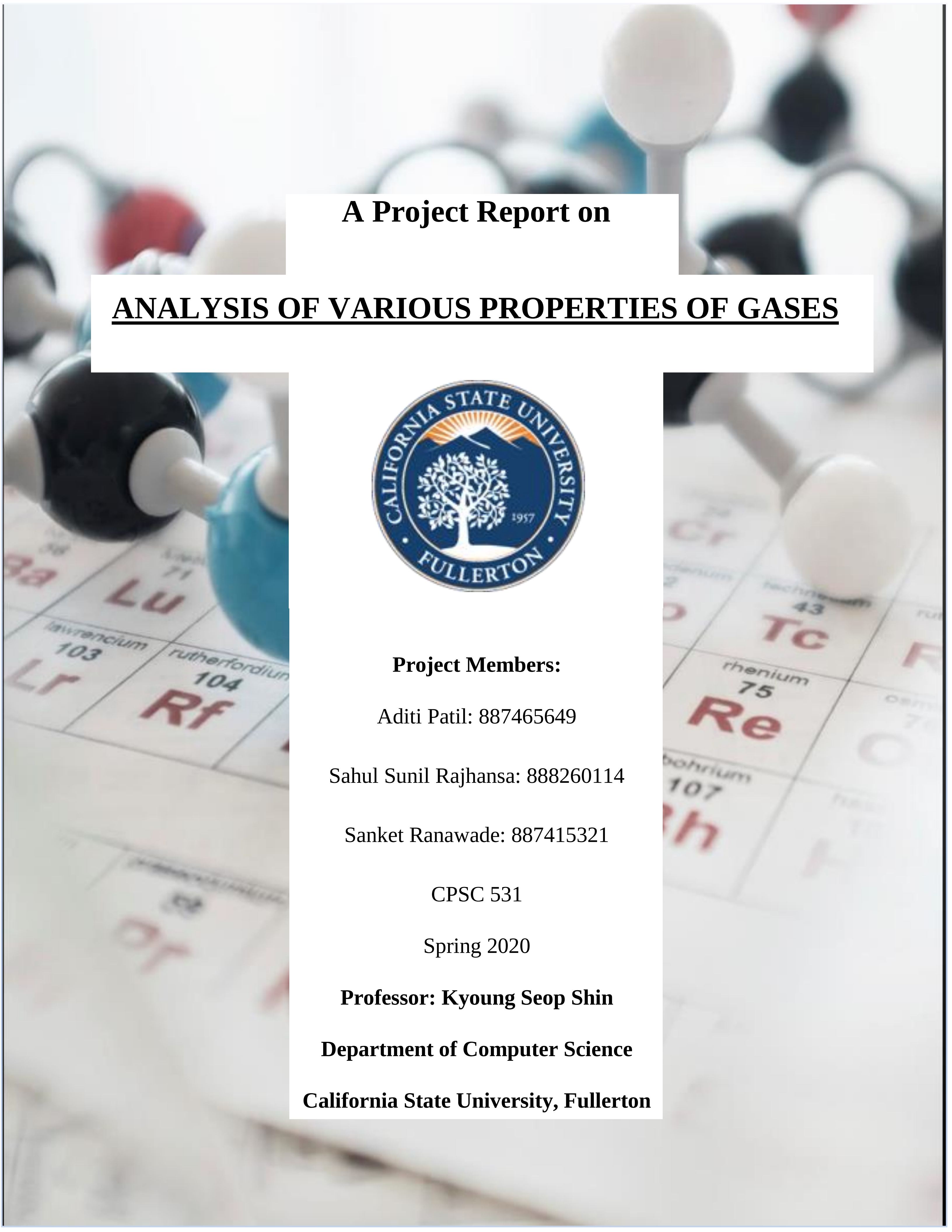


A Project Report on

ANALYSIS OF VARIOUS PROPERTIES OF GASES

Project Members:

Aditi Patil: 887465649

Sahul Sunil Rajhansa: 888260114

Sanket Ranawade: 887415321

CPSC 531

Spring 2020

Professor: Kyoung Seop Shin

Department of Computer Science

California State University, Fullerton

TABLE OF CONTENTS

Table of Contents

ABSTRACT 3
INTRODUCTION 4
LITERATURE REVIEW 5
METHODOLOGY 6
Background 6
Regression 8
Linear Regression: 8
Polynomial Regression 10
Technologies Used: 12
R2 score: 13
Degree of A Polynimial: 14
DISCUSSION 15
RESULTS 16
Code: 16
Output: 20
Heatmap of correlation matrix: 28
Scatter Plot: 29
RESULT COMPARISON 30
Linear Regression Output: 30
Polynomial Regression Output 30
IMPLICATIONS 31
CONCLUSION 33
REFERENCES 34

ABSTRACT

In this paper we report on the a best predictive model to check if there is a functional relationship between the measured physical properties (Tm, Pr, Th, Sv) and a chemical index, Idx by using two methods: the linear and nonlinear method i.e.

- Linear regression
- Polynomial regression.

The proposed predictive model is able to use a csv excel file as the input which consists of all the data that is the data input of the physical properties of gases like Temperature, Pressure, Thermal Conductivity and Sound Velocity and Chemical Property Chemical Index of gas.

With this data input using python we compare the results of each algorithm. In each case we get a r square score i.e. the R2 Score from which we can determine the result of the best predictive model. The more the R2 score closer to one the more is the model best predictive model.

We have used Linear Regression as it is simple to implement and easy to read the results. Polynomial regression was used as Polynomial provides the best approximation of the relationship between the dependent and independent variable. A Broad range of function can be fit under it. Polynomial basically fits a wide range of curvature.

INTRODUCTION

A chemical engineer measured various properties of a gas and created a training dataset. The measured properties include temperature, pressure, thermal conductivity, and sound velocity. Dataset (Tm- Temperature, Pr- Pressure, Th- Thermal Conductivity, Sv- Sound Velocity, Idx- Chemical Index) where all the attributes are type double. The chemical engineer wonders if there are any functional relationships between the measured physical properties (Tm, Pr, Th, Sv) and a chemical index, Idx. We are using two algorithms, Linear regression and Polynomial regression to find this functional relationship. An analysis will be performed on this model and uncertainty of each model will be determined. Depending on the analysis we will determine which is the best method for the process.

LITERATURE REVIEW

The main aim of the project is to check if there is a functional relationship between the physical properties of gases namely: Tm- Temperature, Pr- Pressure, Th- Thermal Conductivity, Sv- Sound Velocity. And Chemical Index (Idx) and develop a best predictive model to determine the relationship.

Using cross validation, we are going to find the best predictive method to find the relationship. For this purpose, we are using two methods: Linear Regression and Polynomial Regression.

Performing tests on a csv MS excel file which contains the data around 2375 rows we will run that data through a code, in which 67% of data will be trained and 33% data will be tested. A predict method will be called.

With the Regressions applied we will get the value for Linear Regression R2 score, For Polynomial Regression we will get the R2 score for each of its degree. Whichever value of R2 score is near to 1 i.e. 100% will prove that method is the best prediction method.

This project includes a limited data i.e. a csv MS excel sheet with limited amount of data input 2375 rows of data is processed.

METHODOLOGY

The main aim of the project is to Develop a best predictive model to check if there is a functional relationship between a chemical index - Idx and the measured physical properties:

- Tm - Temperature
- Pr - Pressure
- Th – Thermal Conductivity
- Sv – Sound Velocity

by using two Algorithms: the linear and nonlinear method:

LINEAR REGRESSION

POLYNOMIAL REGRESSION

Background

Gas Temperature:

The temperature of a gas is a proportion of the normal translational motor vitality of the atoms. In a hot gas, the particles move faster than in an infection gas; the mass proceeds as in the past, anyway the dynamic essentialness, and consequently the temperature, is progressively important considering the extended speed of the molecules.

Thermal conductivity Of Gas:

The most common theoretical explanation of heat conduction in gases is provided by the kinetic gas theory, which treats the collisions between the atoms or molecules as the prime mode of transfer of energy.

Effect of Sound Velocity On Gas:

The speed of sound in a gas is fundamentally a component of the temperature of the gas. Since air is a mix of gases and consolidates water smolder, the general tenacity of air marginally influences the speed of sound. In any case, changes in pneumatic pressure have no veritable effect on the speed.

Gas Pressure:

Pressure is a force exerted by the substance per unit area on another substance. The Pressure of a gas is the force that the gas exerts on the walls of its container. All gases that occupy a space have a pressure of some measurable degree. Gas Pressure is measured in mmHg (millimeters of mercury) with a barometer, which gives the barometric pressure. Pressure is directly measured with a device called a barometer.

Regression

The term regression is used when you try to find the relationship between variables. This relationship is used to predict the outcome of future events.

➢ ***Linear Regression:***

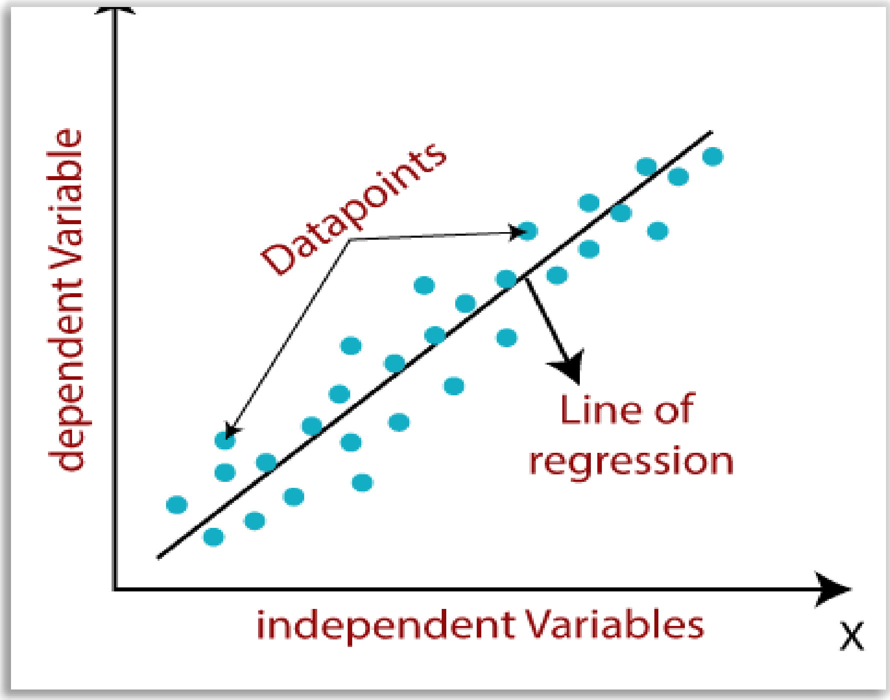
- Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression.
- In statistics, linear regression is an approach to model the relationship between a scalar response and one or more explanatory variables.
- The instance of one illustrative variable is called basic straight relapse.
- For more than one logical variable, the procedure is called numerous direct relapses.
- This term is from multivariate straight relapse, where various related ward factors are anticipated, as opposed to a solitary scalar variable.
- Practical uses of Linear Regression are various. If the goal is prediction, forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables.
- After growing such a model, if extra estimations of the illustrative factors are gathered without a going with reaction esteem, the fitted model can be utilized to make a forecast of the reaction.
- If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in

particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response.

- Mathematical Representation of linear regression:

  y= a0+a1x+ ε

- A general linear regression graph looks like:



- Linear Regression is a way to explain the relationship between a dependent variable and an explanatory variable.
- Linear Regression is a process of finding a line that best fits the data points available on the plots.
- The Independent variables is on X-Axis and the dependent variable is on Y-Axis.

- Data Points are the observations and the straight line drawn which covers maximum number of observations is called as line of regression.
- In our project the Independent Variables are: Tm, Pr, Th, Sv Dependent Variable: Idx

➢ ***Polynomial Regression***

- It is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial. The Polynomial Regression equation is given below:

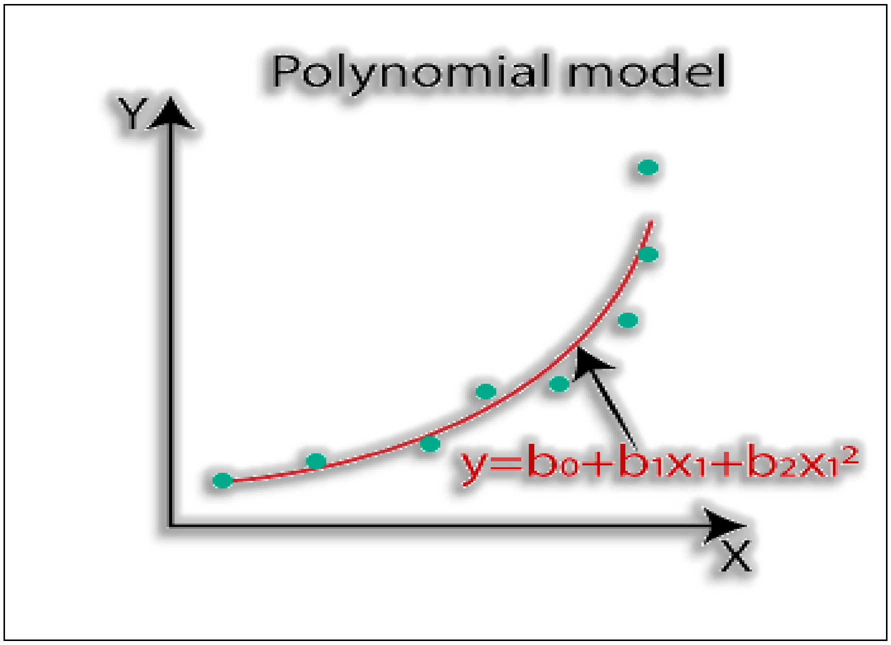
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \varepsilon.$$

Xn gives the Parabolic effect to the graph.

- If your data points clearly will not fit a linear regression, then we use the polynomial regression.

  Polynomial Regression is a form of Linear Regression in which the relationship between the independent variable x and dependent variable y is modeled as an nth degree polynomial.
- In Polynomial Regression, the relationship between the independent variable x and dependent variable y is modeled as an nth degree polynomial. We can do a polynomial regression on the data to fit a polynomial equation to it.
- With polynomial regression, the data is approximated using a polynomial function.

- ***Advantages of using Polynomial Regression:***
  - Polynomial provides the best approximation of the relationship between the dependent and independent variable.
  - A Broad range of function can be fit under it.
  - Polynomial basically fits a wide range of curvature.
- ***Disadvantages of using Polynomial Regression:***
  - The presence of one or two outliers in the data can seriously affect the results of the nonlinear analysis.
  - These are too sensitive to the outliers.
  - In addition, there are unfortunately fewer model validation tools for the detection of outliers in nonlinear regression than there are for linear regression.
  - General Polynomial Regression is represented as:



- Polynomial models have a curve that is an unanticipated turn in inappropriate direction.
- The curve is suitable to cover most of the data points.

Technologies Used:

- IDE: Spyder
- Language: Python
- Algorithms: Linear Regression, Polynomial Regression.
- Scientific Library: Anaconda
- Relative Packages: Pandas, numpy, sklearn, Matplotlib
- Dataset: MS Excel- csv file

R2 score:

- R Squared (the coefficient of determination or R2), tells you how much variation in y is explained by x-variables. The range is 0 to 1, where 0 is 0% variation and 1 is 100% variation.
- R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.
- The definition of R-squared is straight-forward; it is the percentage of the response variable variation that is explained by a linear model. Or:
- R-squared = Explained variation / Total variation
- R-squared is always between 0 and 100%:
- 0% indicates that the model explains none of the variability of the response data around its mean.
- 100% indicates that the model explains all the variability of the response data around its mean.
- In general, the higher the R-squared, the better the model fits your data.
- Theoretically, if a model could explain 100% of the variance, the fitted values would always equal the observed values and, therefore, all the data points would fall on the fitted regression line.

Degree of A Polynimial:

- In mathematics, the degree of a polynomial is the highest of the degrees of the polynomial's monomials (individual terms) with non-zero coefficients.
- The degree of a term is the sum of the exponents of the variables that appear in it, and thus is a non-negative integer.
- The following names are assigned to polynomials according to their degree:

- Special case – zero
- Degree 0 – non-zero constant[4]
- Degree 1 – linear
- Degree 2 – quadratic
- Degree 3 – cubic
- Degree 4 – quartic (or, if all terms have even degree, biquadratic)
- Degree 5 – quintic
- Degree 6 – sextic (or, less commonly, hexic)
- Degree 7 – septic (or, less commonly, heptic)
- Degree 8 – octic
- Degree 9 – nonic
- Degree 10 – decic

DISCUSSION

This project is about finding a fundamental relationship between the physical and chemical properties of gases. As mentioned above we used two regression models to predict which is the optimal model to find the relationship. For this purpose, we are used two methods: Linear Regression and Polynomial Regression.

Performing tests on a csv MS excel file which contains the data of the properties of gases, around 2375 rows we will run that data through a code, in which 67% of data will be trained and 33% data will be tested. A predict method will be called.

With the Regressions applied we will get the value for Linear Regression R2 score, For Polynomial Regression we will get the R2 score for each of its degree. Whichever value of R2 score is near to 1 i.e. 100% will prove that method is the best prediction method. The Result for Linear Regression was R2 score = 0.9783321790408116. The Result for Degree 5 of Polynomial Regression was R2 = **0.9984526964574567**

As we can see from above results the R2 Score of Degree 5 of Polynomial Regression is closer to 1 than that of the linear regression R2 score and any other degree R2 score. Hence, we conclude that the Polynomial Regression model is the best predictive model for analysis of various properties of gases.

This project includes a limited data i.e. a csv MS excel sheet with limited amount of data input 2375 rows of data is processed. We can process databases with more number of rows and columns as we know the efficient and optimal predictive model to analyze the data.

RESULTS

Code:

```
#!/usr/bin/env python3


"""This file creates relational model between different Gas
properties"""


# improting dependencies
import pandas as pd
import numpy as np
from sklearn import linear_model
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score
from sklearn.linear_model import *
from sklearn.preprocessing import PolynomialFeatures
from sklearn.pipeline import make_pipeline
import seaborn as sns
import matplotlib.pyplot as plt
 #%matplotlib inline
 # read data
```

```
df_main = pd.read_csv('Dataset.csv')

# let's peek into the dataset

df_main.head()

# describe the dataset

df_main.describe()

df_main.info()

# calculate the correlation matrix

corr = df_main.corr()

# plot the correlation heatmap

'''sns.heatmap(corr, xticklabels=corr.columns,yticklabels=corr.columns)

plt.scatter(df_main['Idx'], df_main['Sv'])

plt.show(df_main['Idx'], df_main['Sv'])

plt.scatter(df_main['Idx'], df_main['Th'])

plt.show(df_main['Idx'],df_main['Th'])

columns_x = ['Th', 'Sv', 'Tm', ' Pr']

column_label = ['Idx']'''

X = df_main.iloc[:, df_main.columns !="Idx"]

y = df_main.iloc[:,df_main.columns == "Idx"]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=42)

print(X_train.shape)

print(X_test.shape)
```

```
reg = LinearRegression()

reg = reg.fit(X_train, y_train)

y_pred = reg.predict(X_test)

r2_score(y_test, y_pred)

# other way of calculating the R2 score

reg.score(X_test, y_test)

print("weights: ",reg.coef_)

print("bias: ",reg.intercept_)

# split array in k(number of folds) sub arrays

X_folds = np.array_split(X_train, 3)

y_folds = np.array_split(y_train, 3)

scores = list()

models = list()

for k in range(3):

    reg = LinearRegression()

    # We use 'list' to copy, in order to 'pop' later on

    X_train_fold = list(X_folds)

    # pop out kth sub array for testing

    X_test_fold  = X_train_fold.pop(k)

    # concatenate remaining sub arrays for training

    X_train_fold = np.concatenate(X_train_fold)

    # same process for y
    y_train_fold = list(y_folds)

    y_test_fold  = y_train_fold.pop(k)

    y_train_fold = np.concatenate(y_train_fold)

    reg = reg.fit(X_train_fold, y_train_fold)

    scores.append(reg.score(X_test_fold, y_test_fold))

    models.append(reg)

print("Scores")

print(scores)

bestLinear = max(scores)

print("++++++++++++++++++++++++++")

print(bestLinear)

print("----------------Polynomial-------------------------")

# polynomial model

for count, degree in enumerate([2, 3, 4, 5, 6]):
    print("Degree ",degree)
    model = make_pipeline(PolynomialFeatures(degree), LinearRegression())
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    print("R2 score: ",r2_score(y_test, y_pred))
    print("coefficiets: ",model.steps[1][1].coef_)
    print("bias: ",model.steps[1][1].intercept_)
    print("---------------------------------")
   print()
```

Output:

```
<class 'pandas.core.frame.DataFrame'>

RangeIndex: 2376 entries, 0 to 2375

Data columns (total 5 columns):

 #   Column  Non-Null Count  Dtype

---  ------  --------------  -----

 0   Tm      2376 non-null   int64

 1    Pr     2376 non-null   int64

 2   Th      2376 non-null   float64

 3   Sv      2376 non-null   float64

 4   Idx     2376 non-null   float64

dtypes: float64(3), int64(2)

memory usage: 92.9 KB

(1591, 4)

(785, 4)

weights:  [[-5.43785108e-01 -1.60560837e-03  3.61334358e+03  5.55861130e-02]]

bias:  [-77.78381733]

Scores

[0.9783321790408116, 0.9732533362890612, 0.9757403443209168]

Best score obtained by linear Regression

0.9783321790408116
```

```
----------------Polynomial-------------------------

Degree  2

R2 score:  0.9977539113087113

coefficiets:  [[ 0.00000000e+00  1.25260479e+00  2.97853052e-03 -4.85960616e+03

  -7.27623305e-02  5.57667969e-03  1.23345911e-05 -6.76552532e+01

  -2.99241920e-04  6.28141939e-08 -5.01662878e-02 -6.06409351e-06

   1.80569675e+05  4.34181572e-01  1.17246732e-04]]

bias:  [42.67089773]

--------------------------------



Degree  3

R2 score:  0.9983082871799924

coefficiets:  [[-1.50711411e-02  4.73499424e-01  2.12030040e-02 -7.28668237

e+03

  -5.73170069e-01 -1.06117595e-02  1.69132260e-04  8.25591243e+01

  -5.48226362e-03  1.37667552e-06 -8.78601326e-01 -6.58589861e-05

  -1.75177475e+02  2.36710332e+01  9.40405245e-04 -5.34095804e-05

   4.30481316e-07  8.30391733e-01 -1.22355660e-05  1.06478282e-08

  -1.31424548e-03 -4.98723540e-07 -3.46308265e+03 -2.80927155e-03

   9.47320414e-06 -1.58149165e-10 -1.66280007e-04  1.15336070e-08

  -2.68767091e+01  8.44408803e-03 -2.99028853e-07 -1.95979760e+00

   1.10452052e+03 -1.43383461e-01  3.36328721e-06]]
```

```
bias:  [143.55854763]

---------------------------------



Degree  4

R2 score:  0.9984137806008117

coefficiets:  [[-9.73704915e-06 -1.09982639e-02  9.71842627e-03 -2.1567968

3e-02

   9.98459489e-02  9.20484663e-03  4.83790504e-04  1.10981216e-02

  -2.39381203e-03 -2.56187655e-06  5.64621870e-02 -3.37286904e-05

   3.73441120e-06  5.28727079e-02 -9.15478593e-04  9.79241654e-05

   4.59660528e-06 -6.72547533e-01 -4.56854288e-05 -1.19863113e-07

  -4.21529505e-02  1.70661054e-06 -2.25142124e-04  2.65698986e-01

  -3.96816566e-06  4.12577396e-09  1.66042035e-03 -1.76838548e-07

   3.42512756e-03 -1.59096256e-02  1.47213284e-06 -6.52245680e-08

  -1.12675418e-04  3.94033676e-02 -6.39578642e-07  4.76833585e-07

   9.89887991e-09 -3.05164652e-03 -2.77936571e-07 -6.95545523e-10

  -1.39132485e-04  3.46281575e-09 -4.94713092e-02  3.70413482e-03

  -2.52716306e-08  3.44781217e-11  1.65547621e-05 -1.46460914e-09

   1.77567557e-01 -4.99766564e-05  5.86006923e-09 -4.94982157e-05

  -1.52928087e-01 -1.21812016e-03  4.82339526e-08 -2.71524123e-13

  -3.84412942e-07  2.40815566e-11 -1.06332003e-01  1.77800905e-05

  -6.29914759e-10  3.07252723e-04  1.02258628e+00 -1.44465170e-04
```

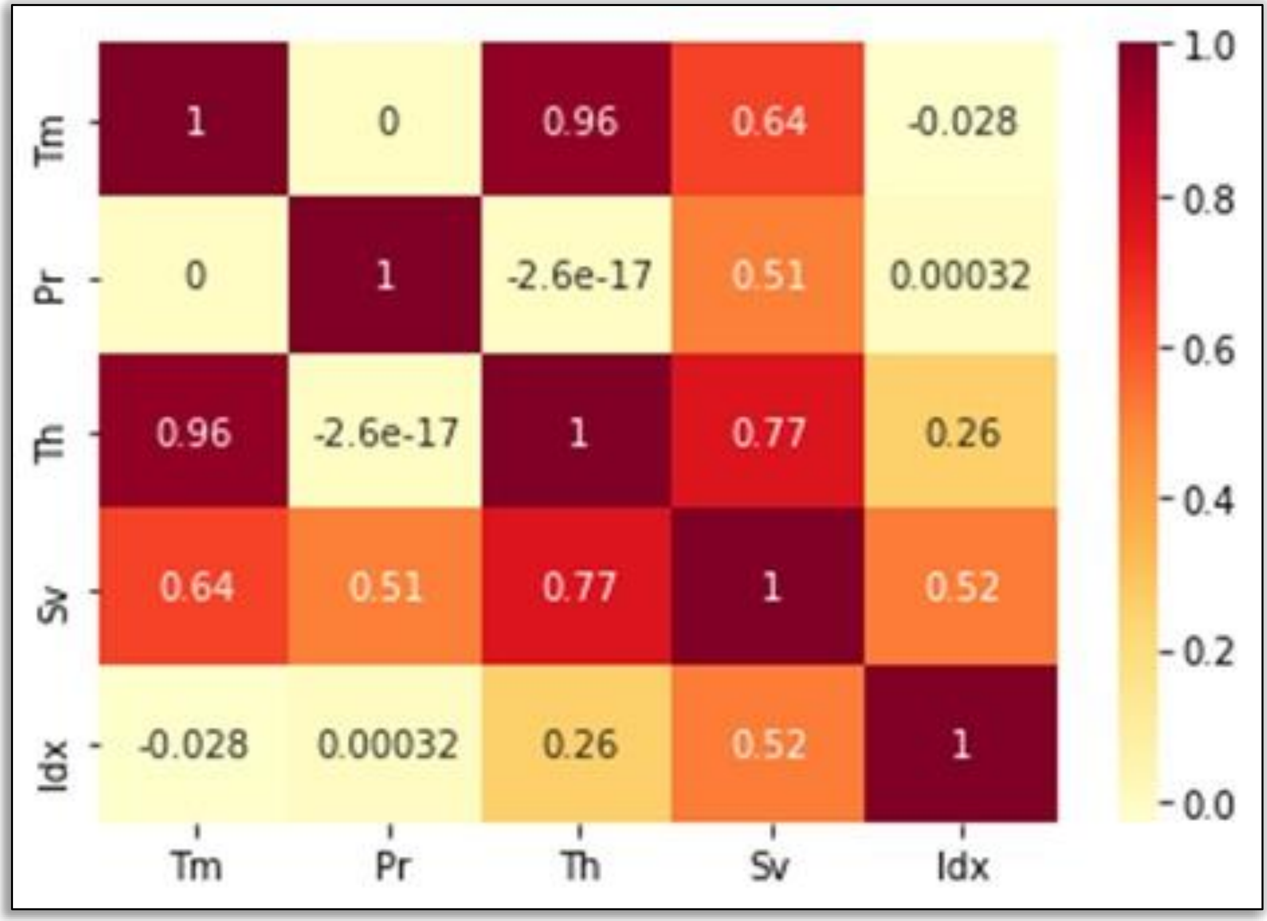
```
   4.55144918e-09 -3.31647685e-08 -1.58666252e-04 -5.13078423e-01
   1.38185960e-04 -7.20588400e-09]]
bias:  [-0.51788081]
--------------------------------


Degree  5
R2 score:  0.9984526964574567
coefficiets:  [[ 1.65743151e-09  8.10019971e-07 -2.46221411e-07 -4.13362947e-07
  -1.50020732e-07 -3.21696562e-06 -2.82937274e-05 -7.99920780e-09
  -1.10100517e-05  1.35969104e-05 -3.68384259e-10 -4.89824929e-05
   5.54586175e-10  3.39879542e-09 -2.09779525e-05  7.79065084e-05
   4.81429839e-07  3.27657311e-08 -2.29968531e-05 -2.45184354e-08
  -9.91517141e-07  1.49652901e-06  9.11769565e-12 -7.89325797e-08
   5.43083719e-06  3.02996872e-09 -5.79806144e-06 -1.27083429e-07
   6.07932299e-10  1.70564031e-07  5.41085853e-07  2.37201277e-13
   4.32133943e-10  1.26037295e-06 -2.46620013e-06 -4.95948636e-07
   1.34796646e-08  3.34735634e-06 -2.44045687e-07 -7.25850592e-10
   9.57990276e-06  1.01266766e-08  3.69839521e-09  9.44493397e-06
   9.11900796e-08 -4.30476559e-11  5.15543666e-06  2.25382933e-10
  -5.90742184e-09 -1.61606189e-04  5.66178578e-09  2.36744275e-12
   7.86412797e-09 -3.03757109e-06 -4.38060776e-08  2.00412816e-12
   7.03258705e-07 -1.03913989e-10  6.26554453e-07 -9.11818562e-06
   1.26293597e-09  6.51126612e-11  2.55555395e-07  1.34349013e-05

  -3.11841553e-09  4.39519500e-15  2.67047049e-11  1.21892918e-07

   2.21373495e-04 -3.58550785e-09 -9.74916745e-09  8.14626942e-11

   6.50768394e-05  3.29552743e-10  2.70014896e-12  5.63435850e-08

  -7.82344880e-11  1.37454160e-07 -2.56694164e-05  1.46712049e-09

  -2.21922307e-13 -3.85005669e-08  5.96298030e-12  2.80390262e-06

   7.97129802e-08 -2.72326410e-11  1.90834263e-10  6.86034094e-07

   3.81830317e-06 -2.55747921e-10  1.30075728e-14  5.47999953e-09

  -4.87981275e-13  6.37036804e-05 -6.116 23265e-08  3.42890001e-12

   2.79500750e-09  1.03050713e-05  2.97565500e-07 -1.39504946e-11

   1.55018871e-13  8.06682272e-10  2.80713320e-06 -9.66904920e-07

   1.06070479e-10 -2.55430832e-16 -2.06721250e-10  1.63912785e-14

  -4.32813637e-05  7.98354699e-09 -2.96824860e-13  1.02635470e-07

   3.84544922e-04 -4.92324194e-08  8.62437766e-13  5.00731442e-12

   2.98399866e-08  1.19611629e-04 -4.45353634e-08  5.00264014e-12

   2.09795882e-16  1.44079874e-12  8.37547571e-09  3.29678933e-05

  -2.44390705e-07  5.64859538e-12]]

bias:  [10.10507652]

--------------------------------



Degree  6

R2 score:  0.9983208330315125
```

```
coefficiets:  [[ 3.39503695e-12  3.33601601e-08  3.40541522e-09  2.92883
082e-09
   2.38247548e-10 -6.61914317e-11 -1.82322111e-10 -1.13761864e-11
  -5.59816067e-11  1.48473181e-09 -4.30581351e-12  2.32017985e-10
   3.26193400e-12  1.81683276e-12 -5.73071637e-11 -1.40661192e-10
  -2.94596276e-09  1.44577530e-15 -9.72053223e-10 -1.02258676e-07
  -2.47672710e-12 -1.98782871e-08  7.23152698e-16  2.88139477e-13
  -3.58442066e-09  2.88327104e-09  3.66763007e-11  2.39623326e-08
   2.69108753e-15  1.64246349e-11 -8.05509585e-12  5.11220893e-16
   2.55707780e-15  8.37200186e-12 -8.19700524e-09 -3.62852850e-09
   4.85483427e-09  3.25257451e-12 -4.67171100e-09 -2.29394503e-09
  -6.95770517e-11  2.14303868e-08  4.96489420e-15  1.98024562e -11
  -2.88966673e-08 -8.92479692e-10 -1.72411518e-09  1.31201930e-08
  -3.07567185e-14 -1.78990764e-10 -4.61555923e-08  6.50713015e-18
   3.74860485e-14  1.69070378e-10  7.54740579e-08 -2.84407292e-11
   6.73556304e-09  5.81882543e-10  9.96309080e-13  4.03671671e-09
  -4.68527648e-09  1.09827096e-16  6.79894275e-13  2.97089221e-09
   1.37240139e-08  6.20483690e-21  7.13110041e-17  4.49599014e-13
   2.06619682e-09 -2.05130544e-08 -5.88275637e-09 -1.11099420e-09
   1.29432644e-10 -2.02457721e-09  1.29115309e-11 -1.67295372e-09
   2.66314587e-10  1.67441053e-13  6.61742654e-10  1.64824260e-09
  -5.50514175e-12 -1.24801493e-07  8.39523741e-11 -2.52035816e-12
```

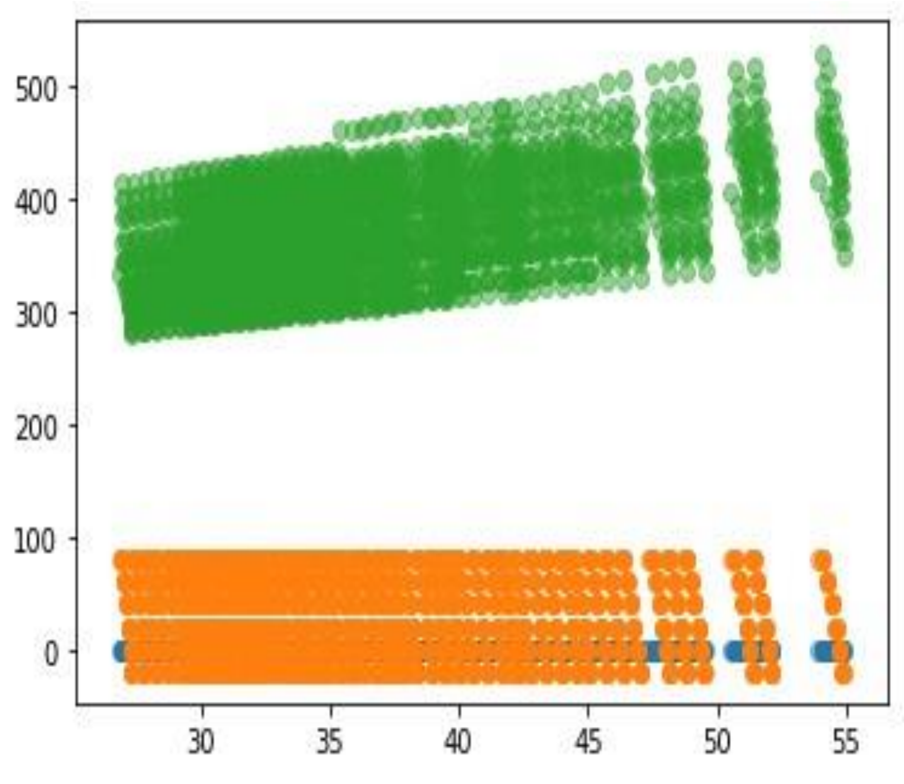
```
-1.21214264e-08 -2.81971864e-10  1.94943167e-16  1.10959015e-12

 4.76080362e-09  9.53736286e-11 -3.30473161e-13  2.99592808e-08

 9.05059308e-12 -6.62580801e-11 -2.79359275e-07 -9.28612295e-11

-6.86870110e-16 -4.77988168e-12 -2.53024303e-08  3.87940319e-10

 2.12659666e-19  1.40310477e-15  8.04033946e-12  3.31753966e-08

-6.22218424e-10 -5.32149579e-15  7.64831218e-10  2.81918726e-13

 2.34509668e-10 -5.33470911e-09 -5.21072089 e-12  2.36698803e-14

 1.00866504e-10 -4.25051208e-08  4.23538464e-11  3.17391133e-18

 2.10535837e-14  1.12004517e-10  3.48552283e-07 -1.39700402e-10

 3.03350454e-22  2.25594165e-18  1.52874785e-14  8.78737233e-11

 3.48209963e-07  1.26586662e-10  2.09880108e-10 -6.24323391e-12

-2.26408006e-10 -1.14100554e-10  4.50860385e-13  1.13253367e-07

-8.12183703e-12 -1.71000320e-12 -1.37574121e-08  4.54800019e-11

-9.38474639e-15 -2.67852466e-09  1.95193575e-13 -2.65551949e-11

-4.81022221e-09  9.78067643e-15 -4.55071604e-16 -1.23689831e-11

-9.81776815e-08 -1.61041433e-12 -1.53459988e-15 -4.46149011e-11

 4.91941775e-14 -2.34585435e-09  1.50696735e-09 -5.23224096e-13

-6.65448917e-14 -2.85759125e-10 -8.29010802e-09  1.94128662e-12

 2.44018674e-18  1.03989824e-14  2.96310194e-11  3.61417221e-08

-3.22306273e-12 -1.84275365e-17  1.25515792e-11  7.54555199e-16

 3.85532108e-08 -2.91336495e-10 -1.38427678e-14 -9.42567388e-13

-1.16940161e-09  1.93290244e-09  1.40495532e-13 -1.89312784e-17
```

```
  -9.11774746e-14 -2.05175665e-10 -2.97000860e-09 -6.81791652e-13

   4.31267914e-21  2.62995839e-17  1.35438104e-13  4.74310877e-10

  -1.09030639e-09  1.29541135e-12  5.30904369e-19  9.60821602e-14

  -1.22781539e-17 -1.97127028e-08 -4.93625282e-12  6.23520006e-17

   3.96406972e-11  1.59208435e-07  8.33812250e-11  1.45561546e-15

   1.27942212e-15  8.97019642e-12  3.97264809e-08 -4.97571518e-10

  -2.55850729e-14  8.52929492e-20  5.99599310e-16  3.54735661e-12

   1.40529916e-08  7.34902600e-10  1.31497 295e-13  7.50600960e-24

   5.44281404e-20  3.57518499e-16  1.96916705e-12  7.27047567e-09

  -1.65573271e-09 -1.11010198e-13]]

 bias:  [9.93117157]
```

Heatmap of correlation matrix:



Scatter Plot:



RESULT COMPARISON

Linear Regression Output:

Scores:

- 0.9783321790408116
- 0.9732533362890612
- 0.9757403443209168

Best score obtained by linear Regression:

- 0.9783321790408116

Polynomial Regression Output

Degree 2 R2 score: 0.9977539113087113

Degree 3 R2 score: 0.9983082871799924

Degree 4 R2 score: 0.9984137806008117

Degree 5 R2 score: 0.9984526964574567

Degree 6 R2 score: 0.9983208330315125

IMPLICATIONS

- We learned what Regression Analysis is Regression analysis mathematically describes the relationship between a set of independent variables and a dependent variable. There are numerous types of regression models that you can use. This choice often depends on the kind of data you have for the dependent variable and the type of model that provides the best fit.
- For this Project we have studied different types algorithms like: Linear Algorithm, Polynomial Algorithm.
- Along with this we studied how to conclude to select the best predictive model. In this process we came across cross validation, hence learned about it.
- To select best predictive model, we studied how to adjust R-squared and Predicted R-squared values.
- We learned that generally, you choose the models that have higher adjusted and predicted R-squared values. These statistics are designed to avoid a key problem with regular R-squared—it increases every time you add a predictor and can trick you into specifying an overly complex model.
- The adjusted R squared increases only if the new term improves the model more than would be expected by chance and it can also decrease with poor quality predictors.

- The predicted R-squared is a form of cross-validation and it can also decrease. Cross-validation determines how well your model generalizes to other data sets by partitioning your data.
- As we have found the best predictive model, we can improve our results by getting in data with large rows and columns.

CONCLUSION

Polynomial Regression is better fit for the dataset as compared to Linear Regression.

Result obtained closely at Degree 5. At Degree 5 the R2 score is closest to 1.

Hence the best predictive method to find the fundamental relationship between the properties of gases is the Polynimial Regression.

REFERENCES

Elmasri, R., & Navathe, S. B. (2015). *Fundamentals of Database Systems (7th Edition)* (7thed.). Hallbergmoos, Germany: Pearson.

O. (2013). 1.3 Physical and Chemical Properties – Chemistry. Retrieved from https://opentextbc.ca/chemistry/chapter/physical-and-chemical-properties/#navigation

Long, J. S., & Freese, J. (1997). Regression Models for Categorical and Limited Dependent Variables. Retrieved from https://www.google.com/books/edition/Regression_Models_for_Categorical_and_Li/CHvSWpAyhdIC?hl=en&gbpv=0

A Refresher on Regression Analysis. (2017, November 30). Retrieved from https://hbr.org/2015/11/a-refresher-on-regression-analysis

Minitab Blog Editor. How to Choose the Best Regression Model. Retrieved from https://blog.minitab.com/blog/how-to-choose-the-best-regression-model

Predictive Modelers' Guide To Choosing The Best Fit Regression Model. Retrieved from https://towardsdatascience.com/predictive-modellers-guide-to-choosing-the-best-fit-regression-model-707120e502b4