Shell Petroleum – Investment in Region according to Prediction on Fuel Type and Car Sales

First name	Last Name	IIT Email	Monday or Tuesday Class	Same project in ITMD 527 (Yes or No)
SAURABH SHEKHAR	PISAL	spisal@hawk.iit.edu	Monday - Wednesday	No

Type of Your Projects (select one or more): Multiple linear regression, ANOVA/two independent Hypothesis testing

1. **Introduction**
   This project is based on the Petroleum company – Shell. The company sells many different types of Fuel and provides the same all over the world. But for this project, we will only be dealing with its reach in the United States. The company has various types of fuel, such as follows – CNG, ELEC, LPG, BD, E85, LNG. Not every vehicle uses all of them and hence the revenue of the company is affected by the productions and sales of these different type of fuels. In this project, we will make models to research how the revenue of east State or Region is affected by the sales of these fuel types. The evaluation will be based on States or Region.
   Secondly, depending upon the sales of the vehicles using these fuel types, the company wants to understand where to invest further which regions to focus more on. We will be covering this part as well.
2. **Data Sets**
   There are 2 data sets
   1. The one which shows data for shell stations across the United States and revenue of each station in according to Fuel type -sales and even other factors contributing to the revenue of each station.
   2. Second data set is of vehicles sold across united states which consume different types of fuels which will help predict future focus of the company.

   The data was collected from the website https://www.afdc.energy.gov/data/
   The First data set for Fuel types has 30,176 rows and Second data set for Car sales has 2162 rows
3. **Research Problems**

   1. By doing a research on the data set, we will be able to get to an evaluation on gas stations in each region/state, which will be affected by the fuel types. By defining the regression, the company will be able to make future predictions and will help the company in decision making such as marketing strategies, offers, fuel deficiency and needs in regions and states, further equipping them with predictions on their transportation of fuel to various locations. It will even help the company decide if they need to open more stations in order to accommodate the requirement/ need of fuel or set up individual stations supplying only 1 particular fuel depending on the way it affects the region's model/states model
   2. The company wants to check for a difference in the mean values of fuel types sold in one particular region for which we will be using independent hypothesis testing or ANOVA to test the difference in means in all Fuel types – this is to understand each fuel has its own contribution towards the company's revenue

4. **Potential Solutions**

   The solution to the problem is to make multiple linear regression models check how the various fuel types are affecting sales in each state/region. The multiple linear regression models will not only contain fuel types affecting the overall revenue of the region but will also contain other factors such as Tips received by customers, food items sold and other supplies which contribute towards the revenue of the station.

5. **Evaluations**

   For evaluation of the solution, I will be using hold-out evaluation and models will be build using all search algorithm such as backward selection, stepwise regression and subset and further it will be evaluated using RMSE for get the best model. (The models will also be tested initially for linearity and residual analysis before testing on test data set)

6. **Expected Outcomes**

   A model with higher accuracy will be created from the past dataset which will help make predictions of the Fuel use and requirements in each region as well as help the company in making decisions on investment. Depending upon the sales of the cars the company will also be able to focus its production and expansion.