Group Member: Chang Li, Qiyuan Xiao, Wenyao Jin

1. What is your dataset?

(1) The size of your dataset

Our size of dataset is 100G.

(2) What are your variables?

For table 'Condition', we select:

sequence_number, plot_sequence_number, state_code, county_code, inventary_year, conditon_status_code, forest_type_code, physiographic_class_code, all_live_stocking_code, unique_condition, carbon_in_soil_organic_material

For table 'Tree', we select:

tree_sequencey_number, plot_sequence_number, tree_inventary_year, tree_state_code, tree_county_code, tree_survey_unit, tree_phase_2_plot, condition_class_number, previous_condition_number, tree_status_code, secious_code, specious_group_name, specious_scientific_name, current_diameter, diameter_height_code_name, diameter_check_code, total_height, actual_height, rotten_and_missing_cull, tree_stocking, mortality_year, unique_tree, crown_class_code, crown_position_code, crown_light_exposure_code, foliage_transparency_code, compacted_crown_ratio.

For table 'Plot', we select:

unique_plot,plot_phase_2_plot_number,plot_state_code,plot_county_code, plot_inventory_year, plot_status_code, sample_kind_code, horizontal_distance_to_improved_road_code, water_on_plot_code, topographic_position_pacific_northwest_research_station, nonforest_plot_status_code, invasive_sampling_status_code, latitude, longitude, elevation

For table 'population',we select:

plot_sequence_numbe,state_code,phase_2_plot_number,land_area_within_the_estimation_unit,total_area_within_the_estimation_unit,land_only,timberland_only,adjustment_factor_for_coarse_woody_debris,adjustment_factor_for_small_fine_woody_debris,adjustment_factor_for_large_fine_woody_debris,start_inventory_year,end_inventory_year

(3) Why did you decide to choose that dataset?

We are interested in American trees development in the past 50 years. And analyzing the data which contains above variables will help us to understand the current situation and past situation about trees in America in the selected years.

The Forest Service has significantly enhanced the FIA program by changing from a periodic survey to an annual survey, by increasing our capacity to analyze and publish data, and by expanding the scope of our data collection to include soil, understory vegetation, tree crown conditions, coarse woody debris, and lichen community composition on a subsample of our plots. The FIA program has

also expanded to include the sampling of urban trees on all land use types in select cities.

2. A visual of your data visualization development

(1) How will you store your data?

The data is stored in the cloud database of kaggle.
https://www.kaggle.com/usforestservice/usfs-fia/home

(2) How will you access your stored data?

We will write code in kaggle notebook. The code is following.

usfs = bq_helper.BigQueryHelper(active_project="bigquery-public-data", dataset_name="usfs_fia")

(3) What languages/frameworks will you use?

We will use Python and Bigquery.

(4) Why did you choose this storage mechanisms, and languages/frameworks?

For storage mechanisms, because the original dataset is too big for us to run in our own computer.

For languages/frameworks, Python is concise and it can connect to Bigquery. Also this is a language which is specialized in data science, data analysis and data visualization.

For Bigquery, we use it because the dataset is a Bigquery dataset, and it is convenient for us to run Bigquery on kaggle.

3. Explain your exploratory data analysis process

(1) What are you trying to highlight from your data?

From table 'Condition', we try to get:

i) Different types of trees in different county or state in a certain time period

ii) carbon in soil organic material, physiographic status, and all live stocking condition change in different county, state in a certain time period leading to the change of type of trees in that area

From table ' Tree', we try to get:

i) tree species summary (where are they, density map of different species in different county, condition change in certain time period, stocking information change,how long they live under certain species in certain county,etc.)

ii) crown situation in certain period under certain species.

iii) what's the situation of trees distribution in next 5 years

iv) what's the relationship between crown situation and mortality

From table 'Population',we try to get:

i)The area of land and timberland in each state

ii)The increment of transect length for different kinds of debris during a period of time

From table ‘Plot’,we try to get:

i) The number of plot which have forest per state

ii) Similar topographic plots in different state

iii)The current geography of the plot（water condition,the distance to improved road,latitude,longitude,elevation）

iv)The situation of invasive plant in plots

(2) What are your overall goals with this visualization?

We are trying to visualize the overall trees development to see if it is consistent with us current forest overall strategy.

(3) Any possible problems/concerns do you have about accomplishing this task

i) It is a really big and a pre-jointed dataset on kaggle without dataset dictionary, and it is different from the original dataset of the FIA program as well. This will cost us a lot of time to read and figure out the instances in the dataset and relationships between tables.

ii) Due to the large quantity of the dataset, we will encounter many dirty data that we need to clean. And it is very likely that after the cleaning, the amount of data will be shrink sharply.

iii) It may take a lot of time for us to find a suitable model to achieve our goal.