You guys can alter the content if you think there is a better way to address the NA values after discussion, if you guys think it's fine then you can keep it. Also, I have already written the code on Google Colab with the change on the below, and you can change the code as well if needed. I will go through this file and coding again after you discuss.

https://www.bloomberg.com/news/features/2020-10-29/small-business-administration-10-000-grant-fraud-went-viral-hurting-program

Missing values

The variables with the missing values: I put my suggestion of addressing the missing values in the ( ).

The below are the variables with the missing values

ppp_applicants_ga_full.csv

1. name: 1 missing observation
   (I think this will not influence our analysis, and we also can delete this column because we can use the variable, Loan_number, as the identifier of each observation. (just like the content in the DataDictionary file says))
2. address: 1 missing observation
   (I think this will also not influence our analysis if we will not do the visualization of this variable. If we want to visualize this, I think we can just filter NA value in the Tableau)
3. naics_code: 2340 missing observations (I think we cannot use any value to replace the NA value in this variable. In class, Professor Adam says that if we want to remove the observation, it is better for us not to remove the percentage of the observations above 3%-5%. 2340 observations are just 0.4% of all data, so I think removing these data is fine.) keep
4. business_type: 41 missing observations (I also think we cannot use any value to replace the NA value in this variable as well, and maybe we can also remove these observations.) keep
5. congressional_district: 5 missing observations (I don't know what is the meaning of this column because this variable is not in the data description. If we want to use this variable, I think we may also need to realize the meaning of this variable; if we don't, I think we can just remove these observations with missing values.) keep
6. undisbursed_amount: 115 missing observations
   (I use 0 to replace the NA value, and the reason is that through the screenshot on the below, we can find that "if the loan status is Exemption 4, the undisbursed_amount will be 0." And I found that this NA value's loan status is also Exemption 4 in the excel file.)
7. project_county_name: 6 missing observations
   (I find the observations with the NA values at first. After then, I use the project_city in the dataset to find the project_county_name, and fill NA with the county name I found in the other observations in the dataset)

8. loan_status_date: 273,239 missing observations (I think this variable doesn't have any NA values because NA value in this variable means that the loan is disbursed but not Paid In Full or Charged Off. (just like the content in the DataDictionary file says))
9. forgiveness_amount: 3,158,142 missing observations (I think the NA values in this variable means that that person doesn't apply for the loan forgiveness, so we don't need to address NA values in this variable) remove
10. forgiveness_date: 3,158,142 missing observations (I think the NA values in this variable means that that person doesn't apply for the loan forgiveness, so we don't need to address NA values in this variable) remove

ppp-removed-ga.xlsx

1. undisbursed_amount: 1 missing observation
   (I use 0 to replace the NA value, and the reason is that through the screenshot on the below, we can find that "if the loan status is Exemption 4, the undisbursed_amount will be 0." And I found that this NA value's loan status is also Exemption 4 in the excel file.)
2. loan_status_date: 2505 missing observations
   (I think this variable doesn't have any NA values because NA value in this variable means that the loan is disbursed but not Paid In Full or Charged Off. (just like the content in the DataDictionary file says))
3. forgiveness_amount: 25836 missing observations remove
   (I think we can just ignore this because all observations' value are NA)
4. forgiveness_date: 25836 missing observations remove
   (I think we can just ignore this because all observations' value are NA)

Type of the variables

I found that some variables have wrong data types, and I just made some changes below. I have already checked the data type of them with the DataDictionary file.

ppp_applicants_ga_full.csv

Numerical to Categorical data type:

naics_code, loan_number, sba_office_code, servicing_lender_location_id, originating_lender_location_id

ppp-removed-ga.xlsx

Numerical to Categorical data type:

naics_code, loan_number, sba_office_code, servicing_lender_location_id, originating_lender_location_id, forgiveness_date

Next step (From my perspective)

I will do tonight: merge the datasets, and create a new column “removed” to identify whether the observation is fraud or not, 1 as fraud; 0 as no fraud.

Finished before discussion on Tuesday: I think maybe we can use data visualization to find the different distributions in variables between two datasets in Python. Then, put two datasets into the Tableau separately to better visualize, especially for the geographical variables. Maybe we also can use a dashboard to do the presentation. (We can just separate variables to each person equally, which is easier for everyone, but we also have to think about how to separate.)

Finished before discussion on Wednesday: After finishing those, we can start to build the regression model to do the prediction. Before doing the logistic regression.

And, I also write my suggestive schedule in the check lists documents, you can change anything if needed.

Wendy:

1. I think when we use the “fraud” variable, we have to be careful since the reason that the application was removed from the dataset is unclear. There might be a great chance that the removal is because of fraud, but right now we only know that the data was removed for some reason.
2. I think we can consider a new variable, “amount/ jobs retained”.
   I am not sure how that will impact the result, but I think this can be interesting. On the other hand, the amount and jobs retained are correlated with each other, so we can consider replacing those with this new one.
3. Thank you Zoe! I think the logic for those NA values all make sense to me!

Chido:

1. Distribution of loans to different counties did that contribute to the removed loans?

amount initial_approval_amount term current_approval_amount undisbursed_amount jobs_retained -Emma

naics_code (hard to recognize)
business_type (similar)
date_approved (No difference: both have the highest in Q2, lowest in Q1 and Q3)
lender (hard to recognize)
Sba_office_code (No difference)
Processing_method (No difference)
Loan_status (Difference)
- Zoe (I have already put the whole plots in the Tableau file on the Google drive)

servicing_lender_location_id servicing_lender_name rural_urban_indicator
hubzone_indicator business_age_description loan_status_date
originating_lender_location_id lmi_indicator, amount/jobs retained - Wendy

Cleansed data for tableau-

state address city zip
servicing_lender_address
servicing_lender_city
Servicing_lender_state
servicing_lender_zip
project_city
project_county_name
project_state project_zip
Originating_lender_city
originating_lender_state
congressional_district - Chido

2. Write down the count per county/ loan amount
3. SBA zones
4. Range of money

Label	Fulton County, Georgia
Total:	1,066,710
Population of one race:	995,802
White alone	418,700
Black or African American alone	453,834
American Indian and Alaska Native alone	3,255
Asian alone	80,949
Native Hawaiian and Other Pacific Islander alone	452
Some Other Race alone	38,612
Population of two or more races:	70,908

Label	DeKalb County, Georgia
Total:	764,382
Population of one race:	714,489
White alone	225,752
Black or African American alone	388,963
American Indian and Alaska Native alone	4,412
Asian alone	50,384
Native Hawaiian and Other Pacific Islander alone	250
Some Other Race alone	44,728
Population of two or more races:	49,893

https://data.census.gov/cedsci/table?q=delkalb%20ga

- Sole proprietors, independent contractors, and self-employed persons
- Any small business concern that meets SBA's size standards (either the industry size standard or the alternative size standard)
- Any business, 501(c)(3) non-profit organization, 501(c)(19) veterans organization, or tribal business concern (sec. 31(b)(2)(C) of the Small Business Act) with the greater of:
    - 500 employees, or
    - That meets the SBA industry size standard if more than 500
- Any business with a NAICS code that begins with 72 (Accommodations and Food Services) that has more than one physical location and employs less than 500 per location

A lot of 812112 NAICS code businesses (greater percentage) - removed, why ? characteristics

Variables we didn't see significant difference between "removed" and "not removed":

1. rural_urban_indicator
2. lmi_indicator

The Active Un-disbursed status means **you were approved for the PPP loan, have signed your contracts, and are waiting for your funds to be dispursed.**

The Active Un-disbursed status means you were approved for the PPP loan, have signed your contracts, and are waiting for your funds to be dispursed

Questions:

1. Definition of loan status

2.