**Option# 2**

Running SAS Code on Dataset Provided by the U.S. Department of Health & Human Services

The health care industry produces advantages and services to treat patients. The industry consists of a variety of markets such as physicians, nurses, insurance, hospital, and pharmaceutical. Crucial data covers the quality, intensity, quantity, and costs of services and come from both patients and providers and in a variety of forms like structured and unstructured.

The Department of Health and Mental Hygiene (DOHMH), NYC Open Data website has a massive set of data that can be used to practice SAS software.

**Popular Baby Names by Sex and Ethnic Group**

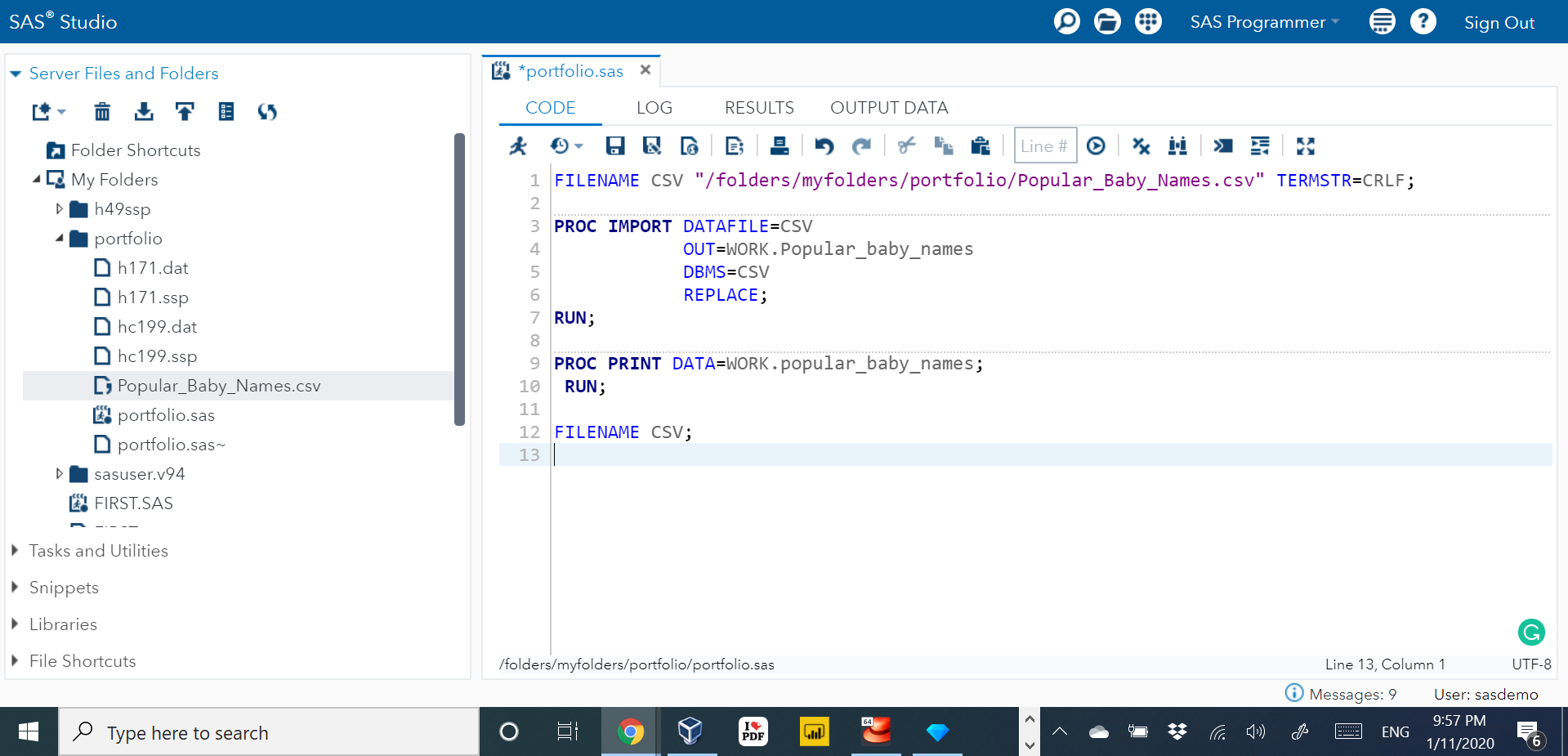
Data were collected through civil birth registration. Each record represents the ranking of a baby name in the order of frequency. Data can be used to represent the popularity of a name.

The CSV file I downloaded from the website has unstructured data. The Ethnicity column has one ethnicity entered twice with different spelling. I utilized Openrefine software to consolidate these categories by using Text Facet option.

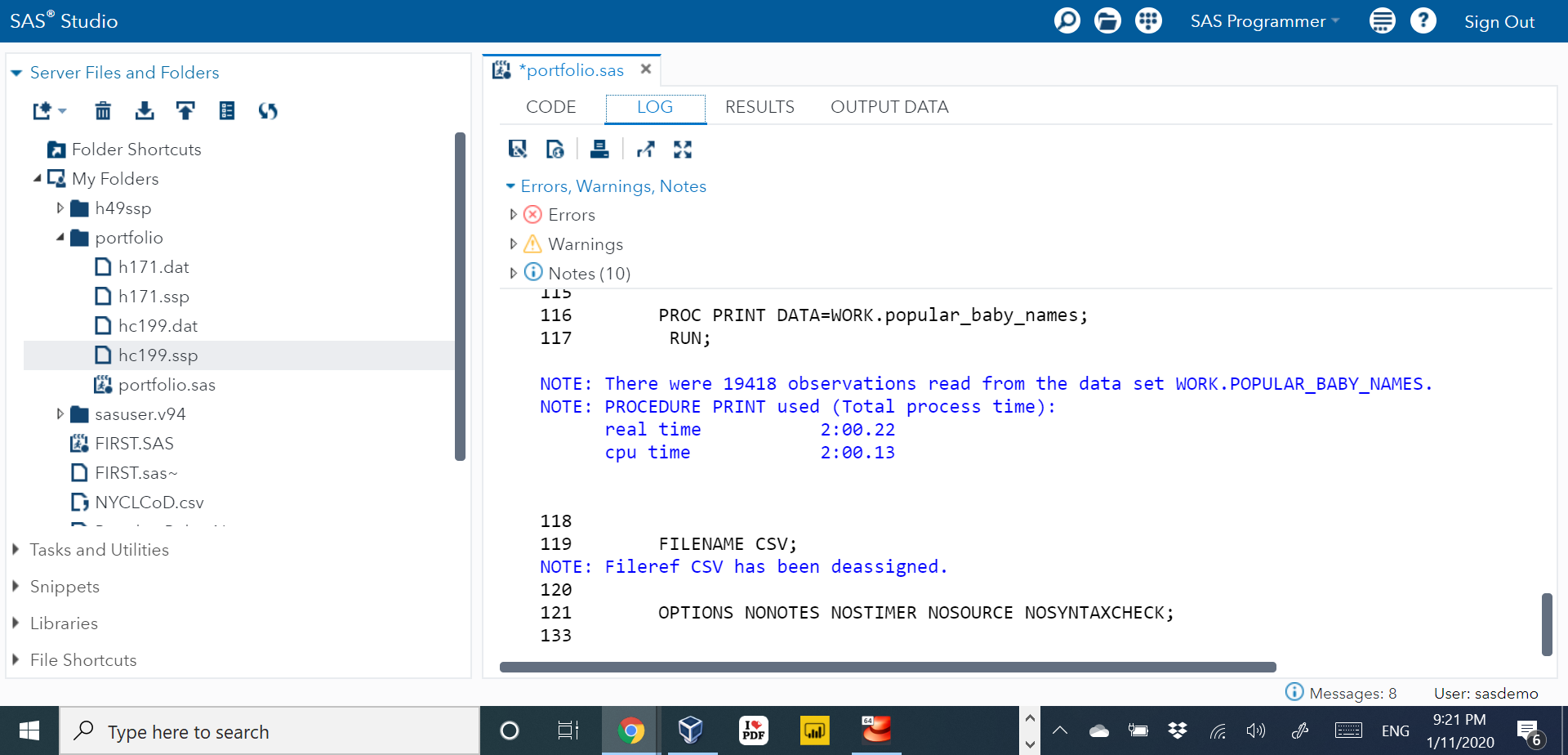
**Before After**



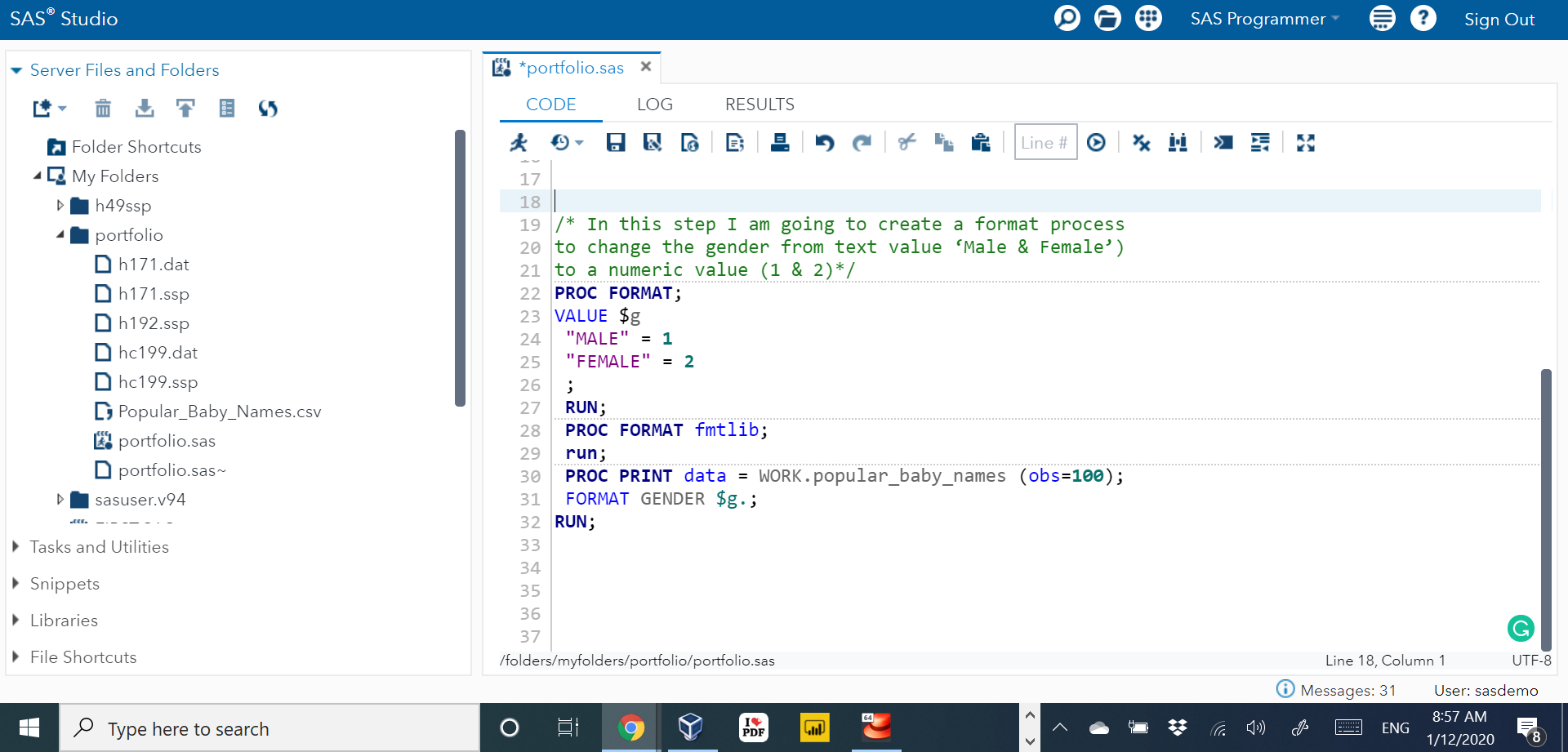
At the beginning, I am going to upload my data which is in csv format into SAS university Edition Software as shown below:

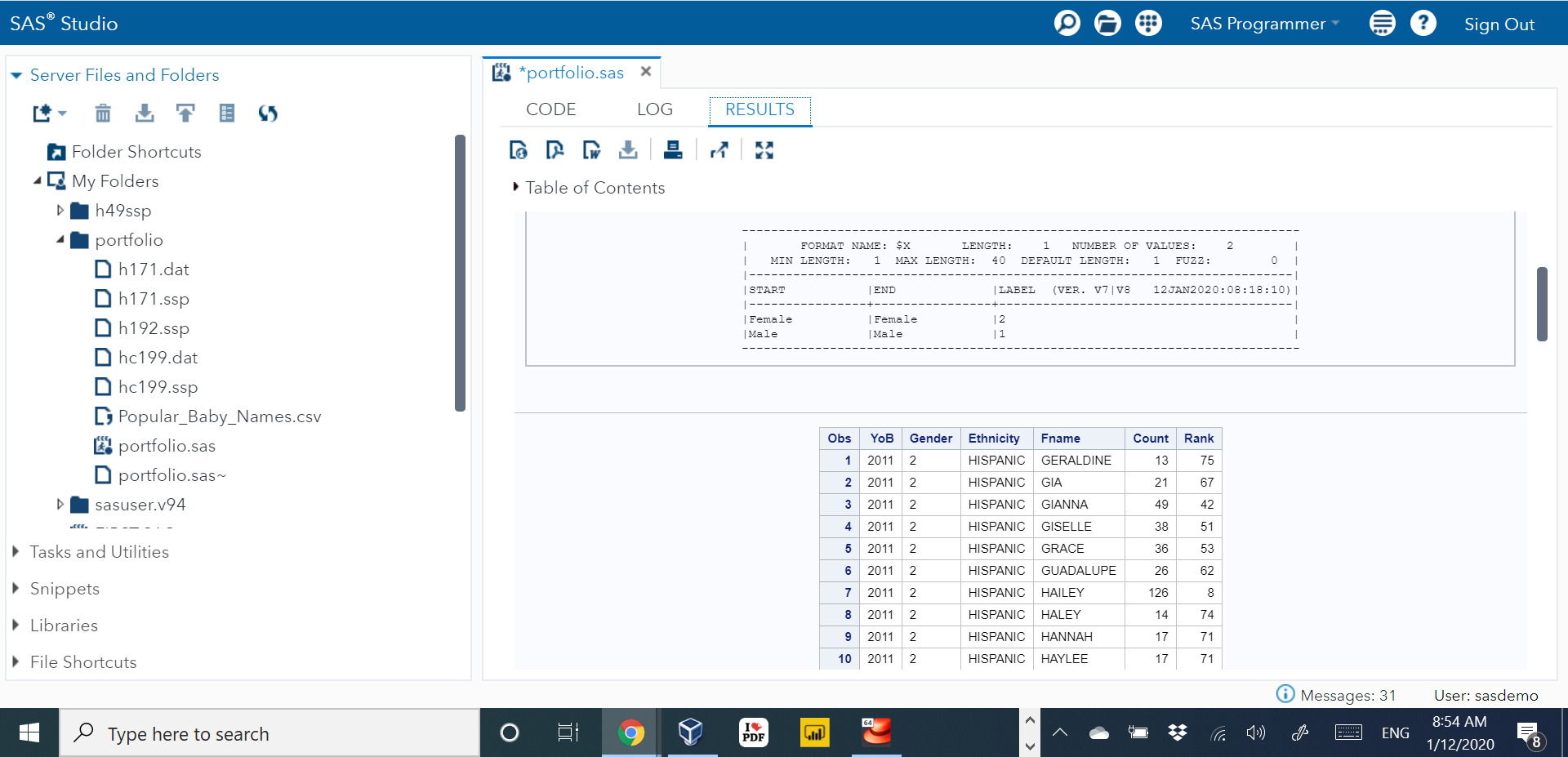


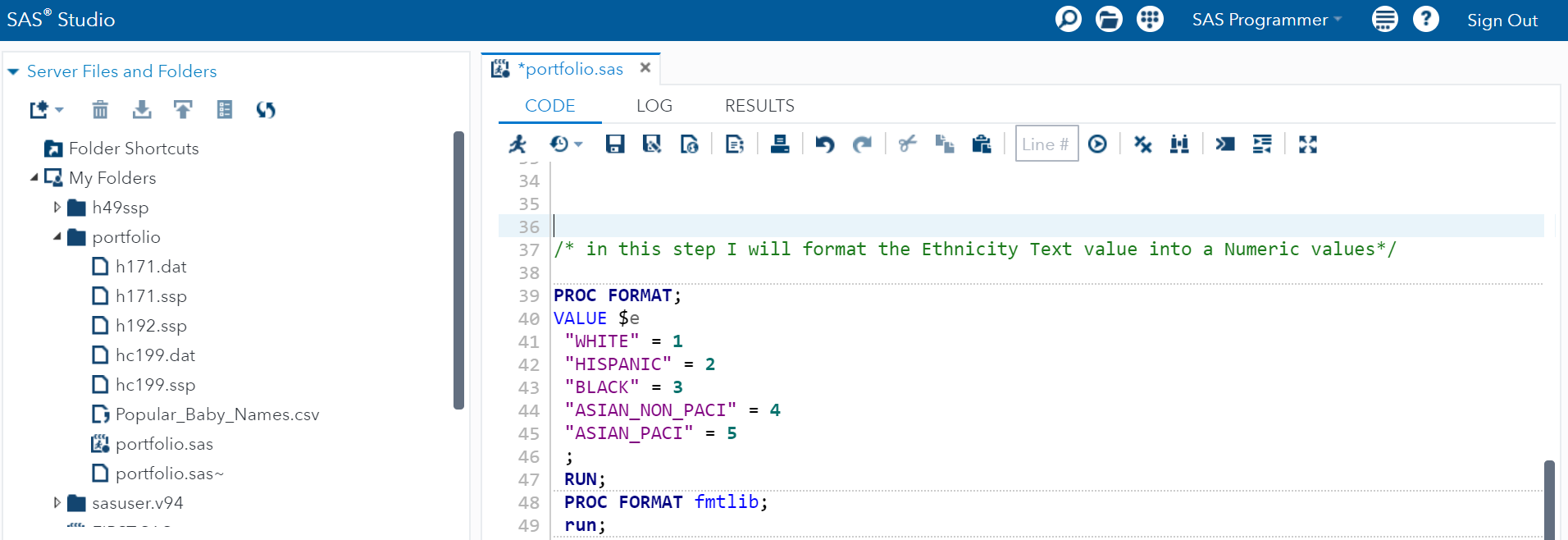
The code ran successfully with no errors as displayed below:

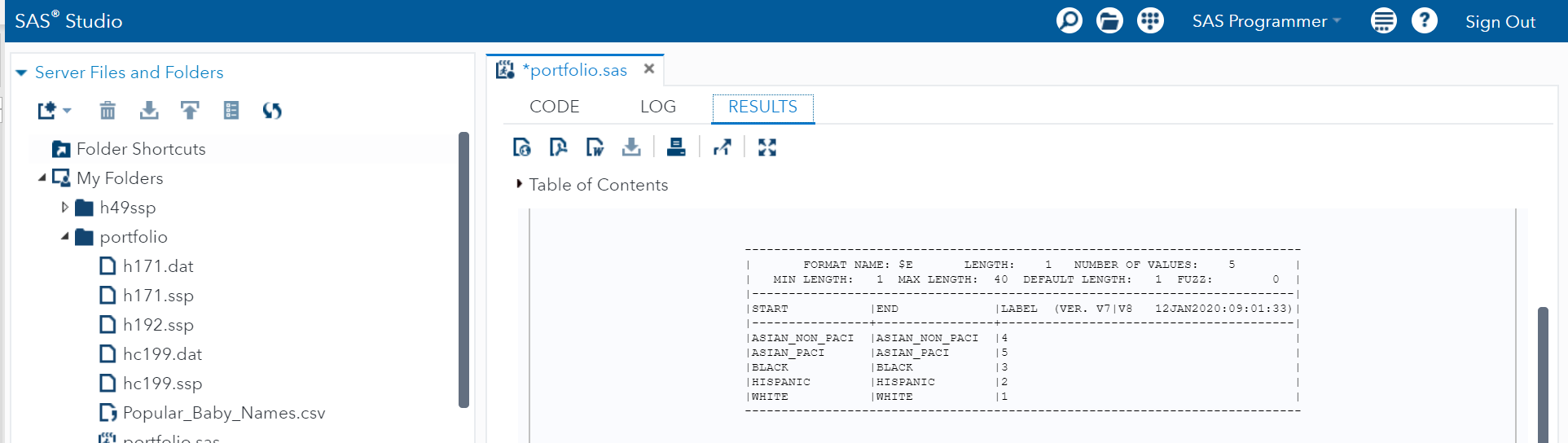


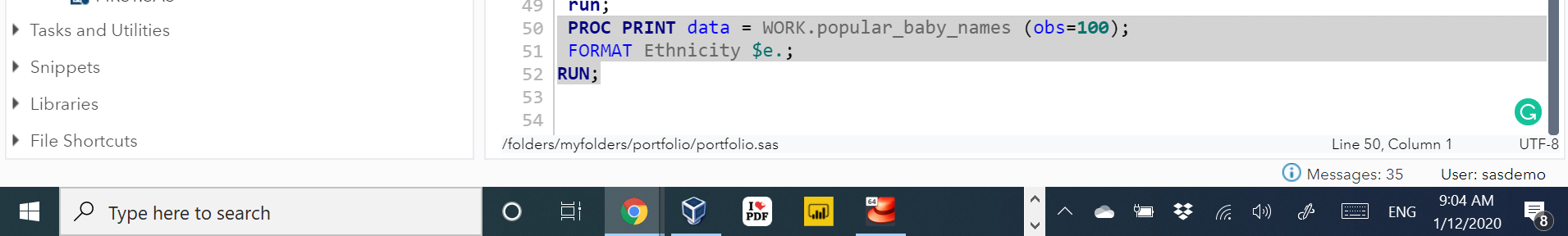
In this step I am going to create a format process to change the gender from text value ‘Male & Female’) to a numeric value (1 & 2)

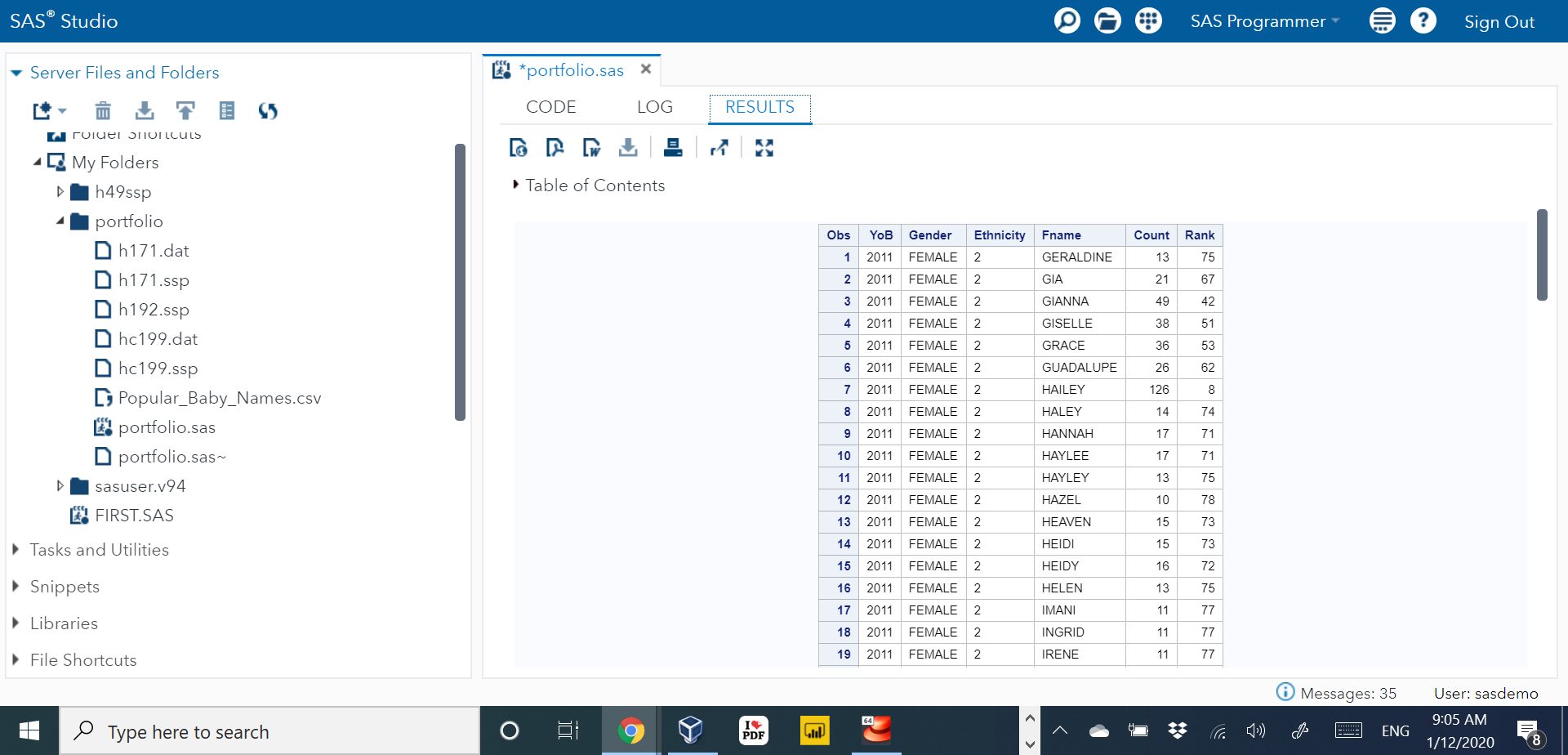






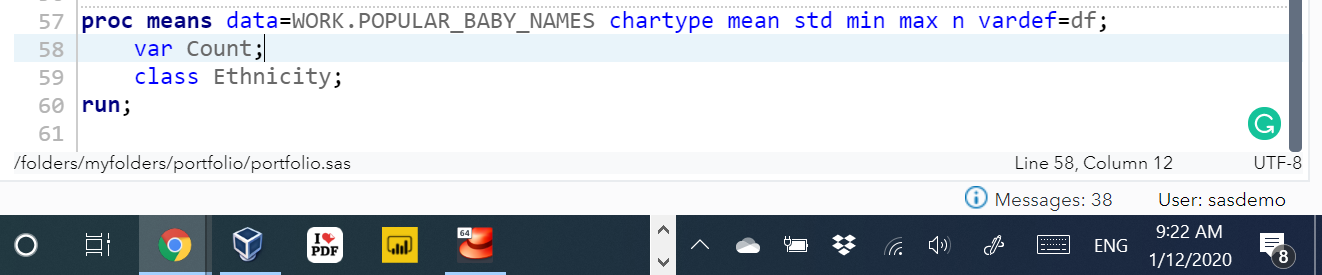


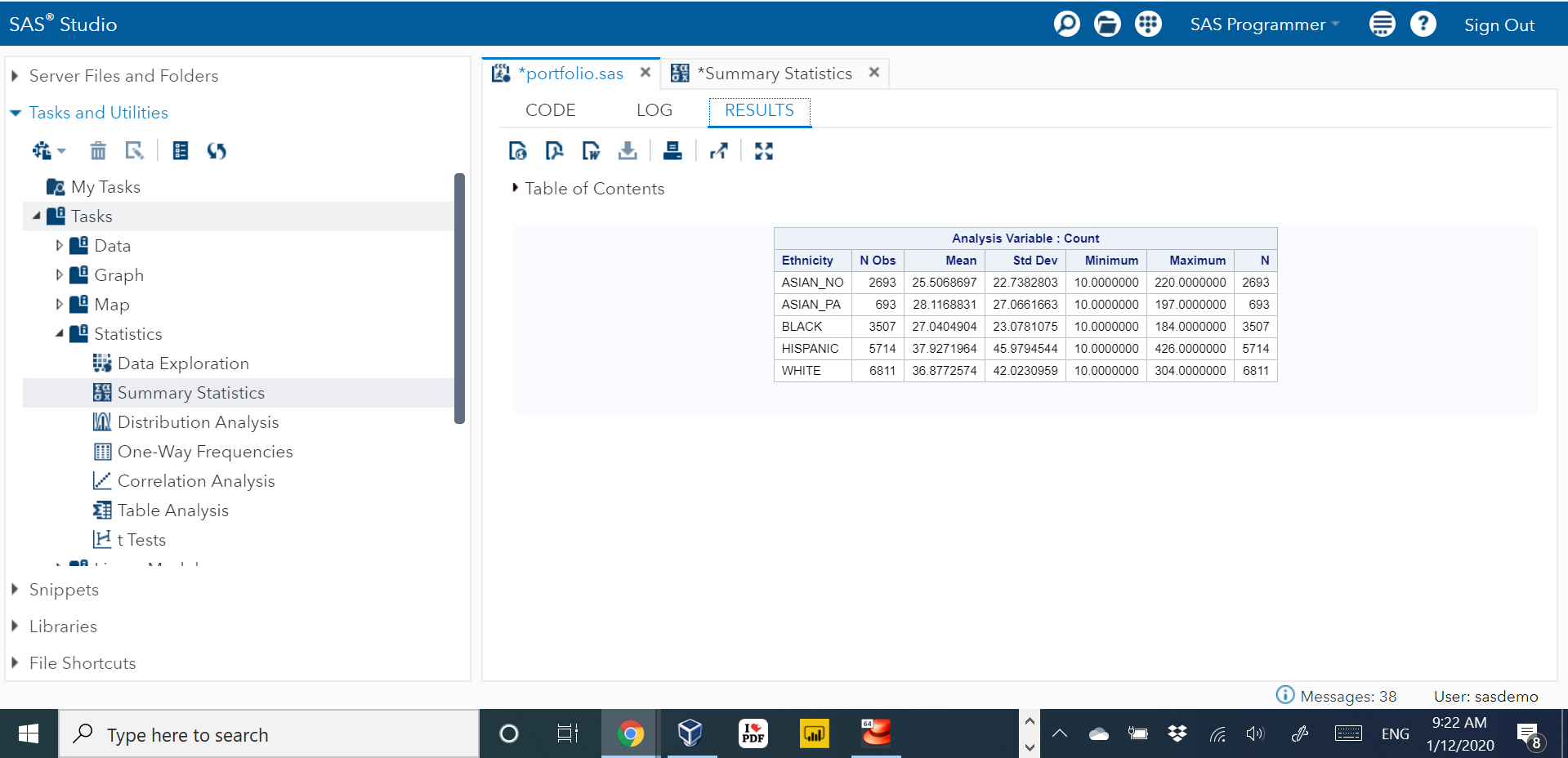




/\* This step is to create a Summary Statistics for the Variable Count

Classified by Ethnicity \*/





\* In this project, I want to test a two-sided hypothesis

to assess whether the mean Count number from males and females are equal.

This test implies a two-sided, two-sample t-test

of whether the difference in the mean is equal to zero \*/

proc univariate data=WORK.POPULAR\_BABY\_NAMES normal mu0=0;

ods select TestsForNormality;

class Gender;

var Count;

run;

/\* t test \*/

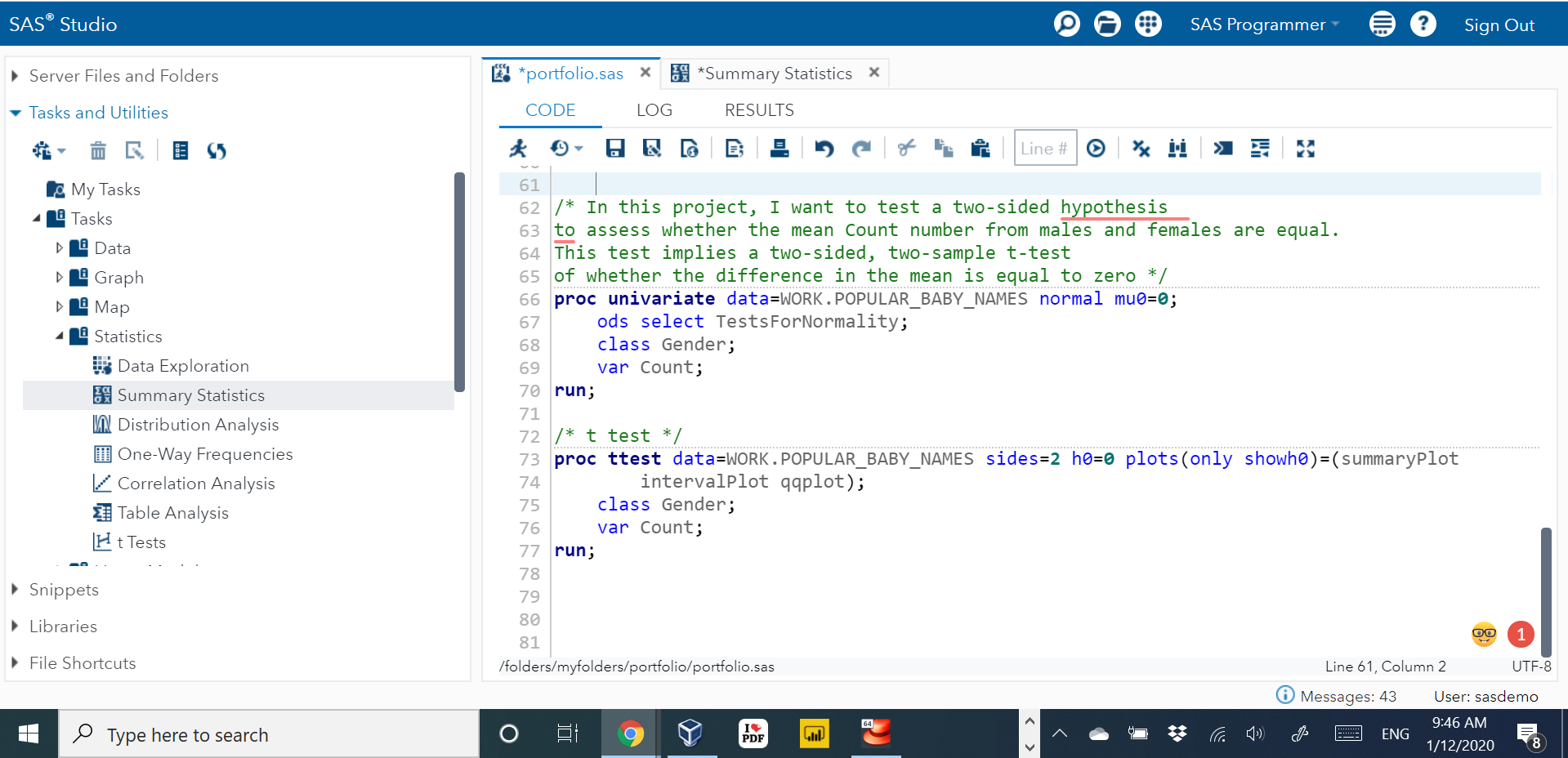
proc ttest data=WORK.POPULAR\_BABY\_NAMES sides=2 h0=0 plots(only showh0)=(summaryPlot

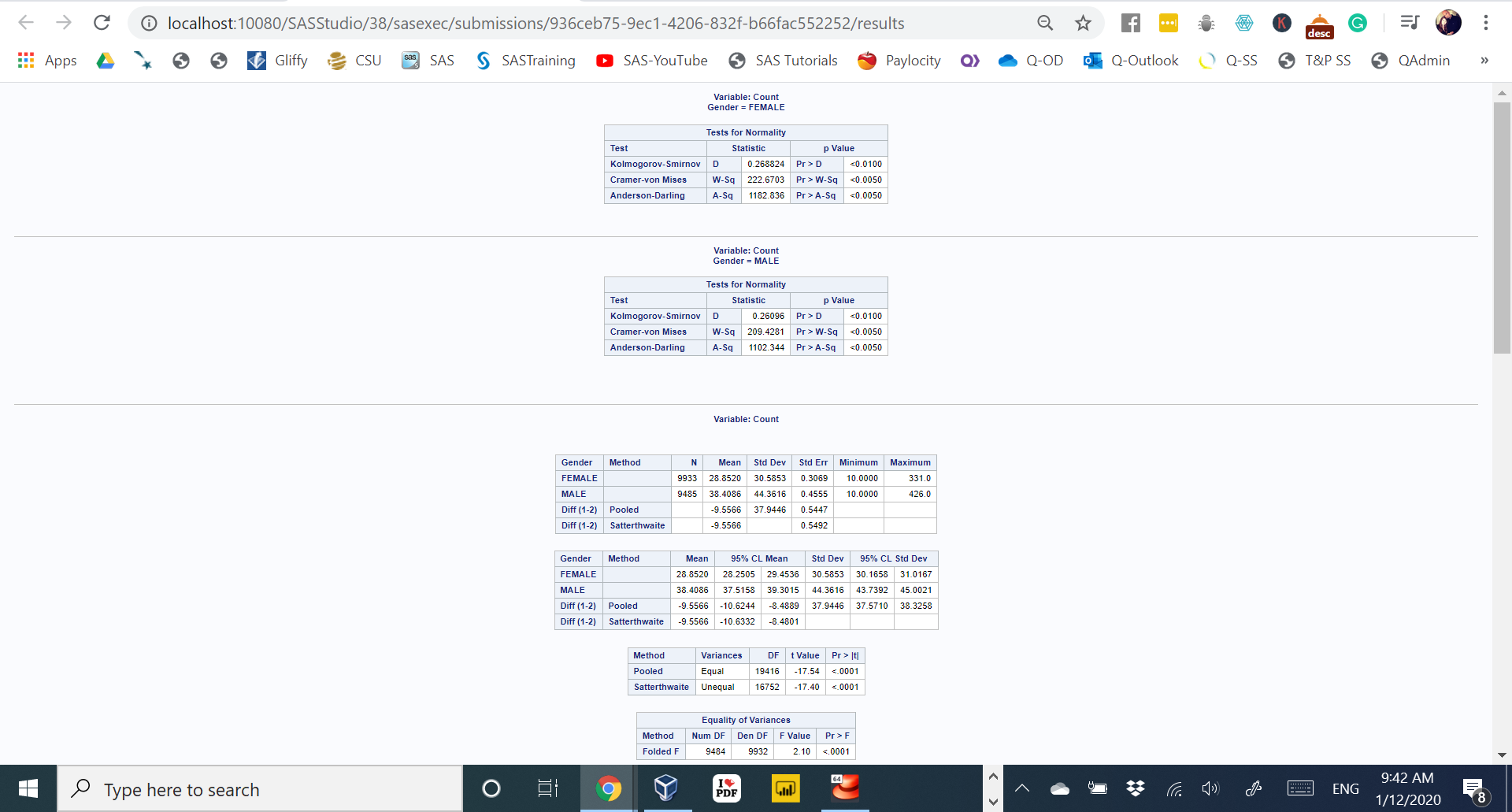
intervalPlot qqplot);

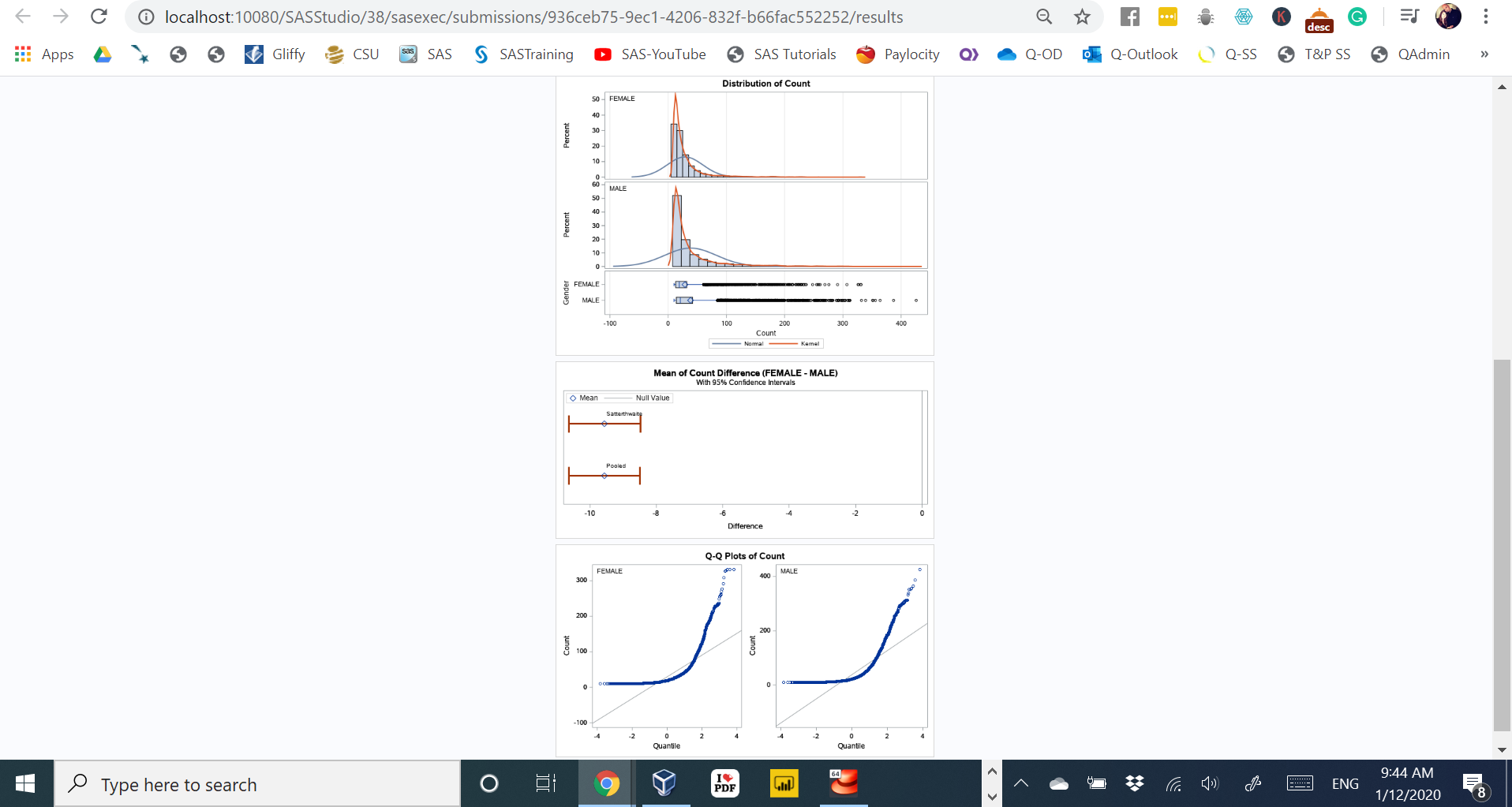
class Gender;

var Count;

run;







By looking at SAS code result, we find out the p-Value = 0.0001 which is marginally lower than 0.05 and that means that there is a significant difference in the average count of names between Males and Females.

**References**

Allison, R. (2017). *Visualizing data with SAS.* Cary, NC: SAS Institute Inc. Retrieved From:

<http://support.sas.com/publishing/vds.pdf>

Carpenter, A. L. (nd). *Simple tests of hypotheses for the non-statistician: What they are and why*

*they can go bad.* California Occidental Consultants, SAS Institute Inc., Vista, CA.

Elliott, A. C., & Woodward, W. A. (2016). *SAS essentials mastering SAS for data analytics (2nd*

*ed.).* John Wiley & Sons, Inc. Hoboken, New Jersey.

*Getting Started with SAS Studio.* SAS Tutorial Video. Retrieved From:

[https://video.sas.com/category/videos/how-to-tutorials#category/videos/how-to- tutorials/0](https://video.sas.com/category/videos/how-to-tutorials#category/videos/how-to-tutorials/0)

Popular Baby Names (2019, June 8). Department of Health and Mental Hygiene (DOHMH): NYC Open Data. Retrieved from: [https://data.cityofnewyork.us/Health/Popular-Baby- Names/25th-nujf](https://data.cityofnewyork.us/Health/Popular-Baby-Names/25th-nujf)

SAS Documentation Support Documents. Retrieved From: <https://support.sas.com/en/documentation.html>