

TAQ Database Operations

Connecting to TAQ and Executing Holden and DowLoop Codes.



May 1, 2018

**The document explains the following points:**

1. **Installing Putty and WINSCP and navigating tips.**
2. **Viewing TAQ datasets through WINSCP and Putty.**
3. **Viewing SAS Programs in PC-SAS and executing the same.**
4. **Moving SAS datasets/Other files from WRDS site to Local machine.**
5. **Viewing SAS datasets in PC-SAS.**
6. **Execution and changes that need to be done to Holden and DowLoop codes.**

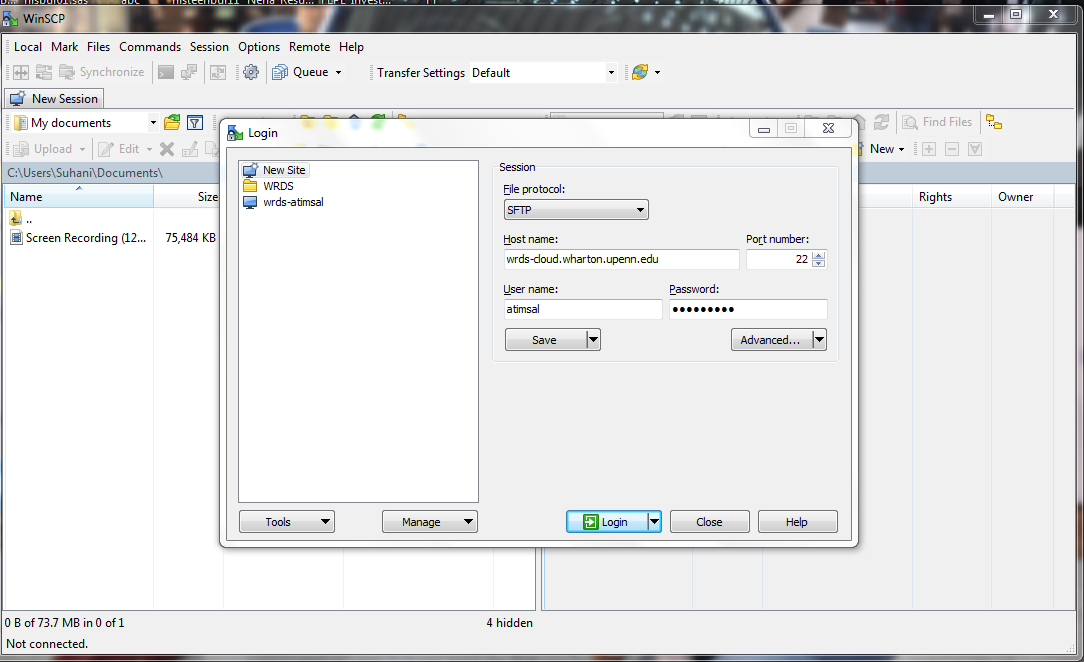
**Task Objectives:** To come up with detailed understanding and complete plan to execute all the codes shared with you already and replicate those codes at my end. Gaining understanding of accessing WRDS though SSH client and dealing/working with datasets.

To achieve the above Task Objectives, I have formulated various steps:

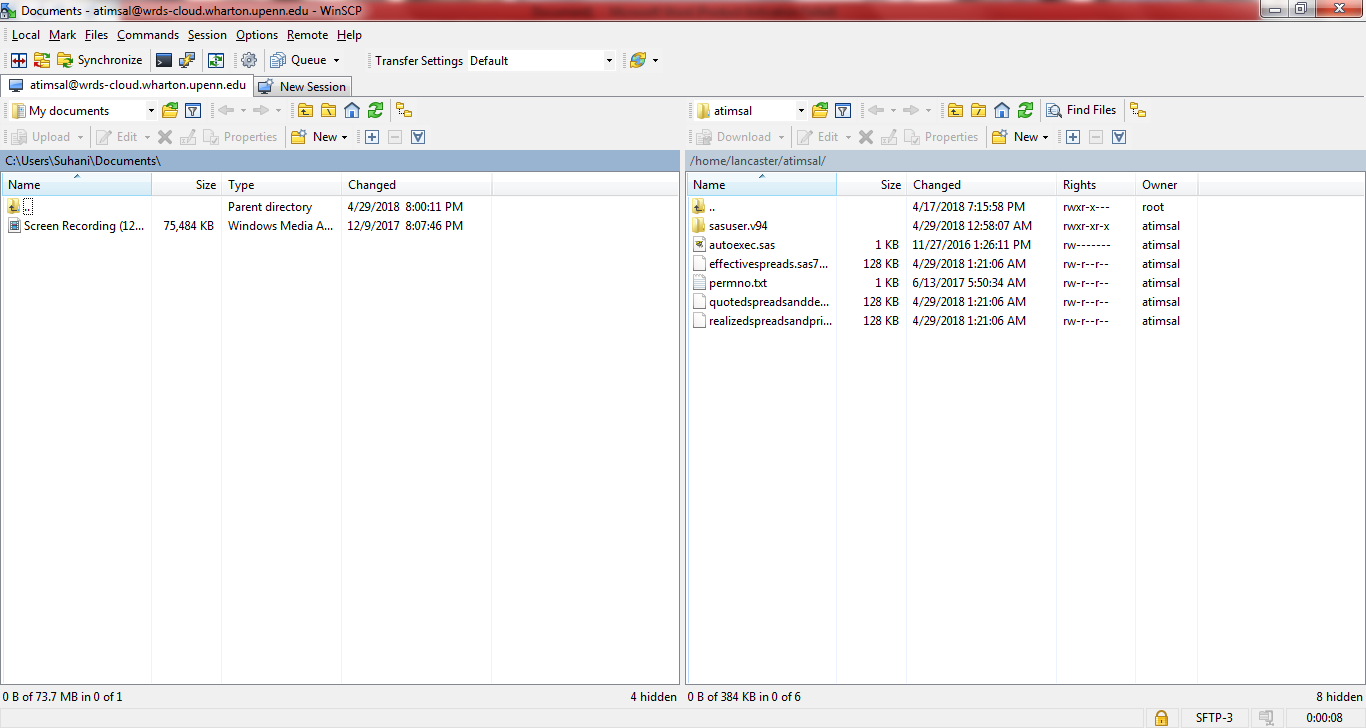
**Step1:**

To work on the TAQ datasets we need to setup the software platform and understand the platform how it works. Below are the software applications that needs to be available/installed on your PC to work and run the TAQ codes.

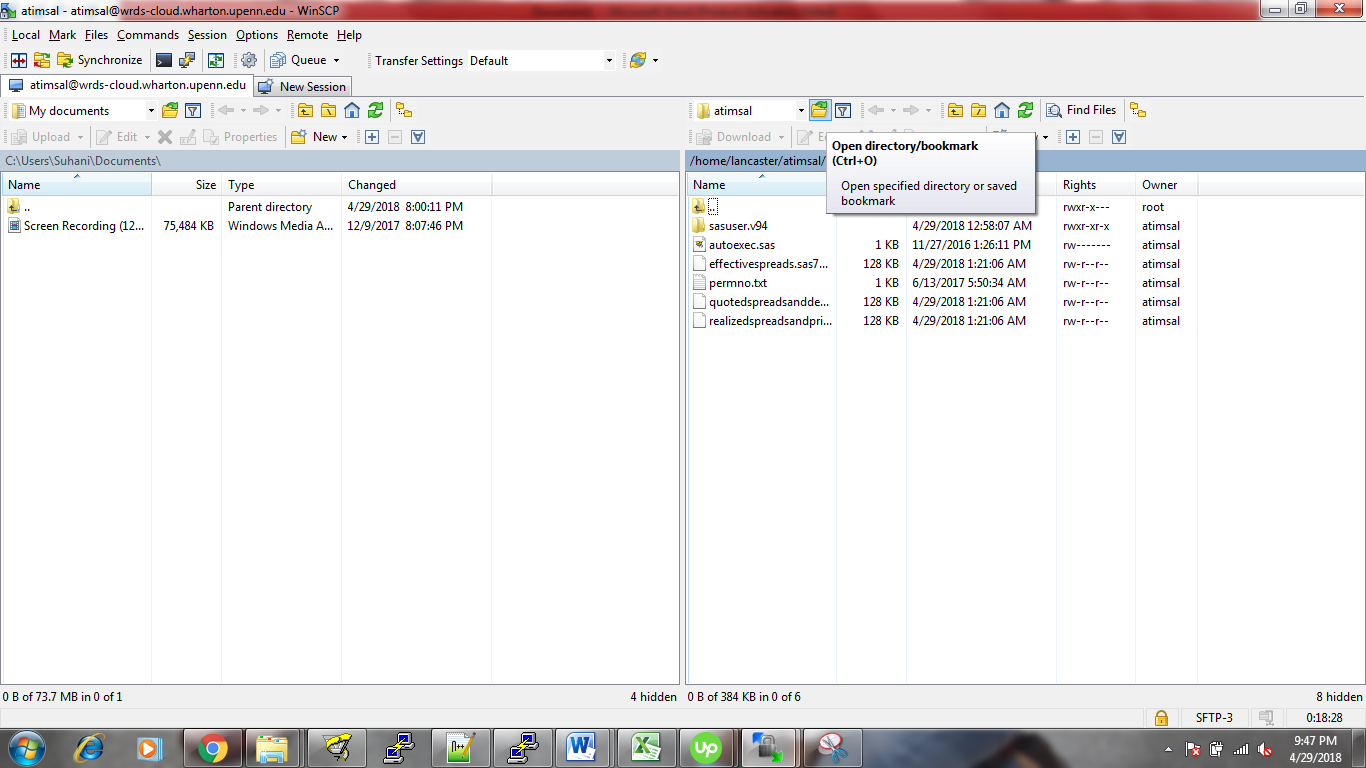
* WINSCP 🡺 It’s a windows based software application used to view the datasets/files present on Linux/Unix server system(WRDS is installed on Linux OS). We can download WINSCP from the link <https://winscp.net/eng/download.php>. Click the download button and after clicking the download button it will ask for regular installation parameters of software and installs the same. After installing WINSCP open it and select New Site in the left side area after that on the right side of window hostname field give **wrds-cloud.wharton.upenn.edu** and port number should be default as 22. And in the user name give your WRDS user name and password in the password field. As shown below:

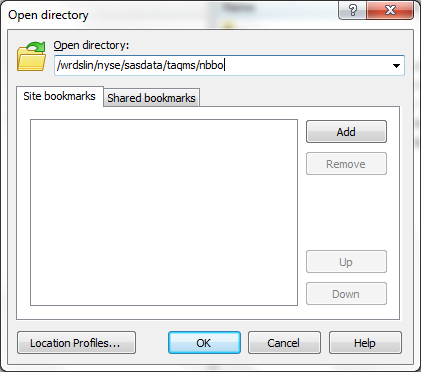


Once you provide details there is an option to save those parameters else you can directly login. Once you click login it will show as below:



So in the above image window left side area is your local PC and right side window is the WRDS system and it will directly land into the home directory so in this case its **/home/lancaster/atimsal**. So this will also provide you access so that you can directly move files from your local PC to WRDS Linux system vice-versa. This can be done directly by dragging from left side of the pane to right side or from right side to left side(For copying files from WRDS server to local PC). If you want to view the files from TAQ database you need to open the directory by clicking on the icon below in the tool.

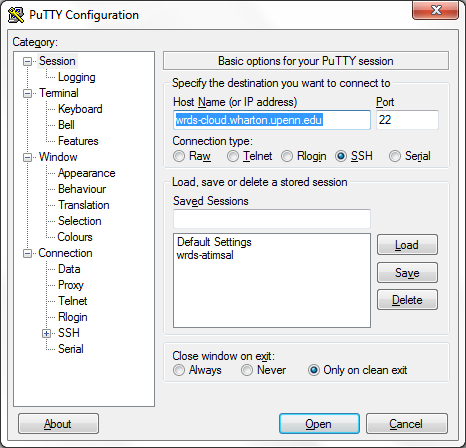
 After clicking on the icon as shown above you can type the directory as /wrdslin/nyse/sasdata/taqms/nbbo and click OK to view the nbbo files.( Not recommended though as it takes lot of time for me to view the files as there are a lot and of big size). In this way you can always naviagte to other directory and view the files.



* Putty 🡺 It is used for logging into the WRDS server in a secured way. You can download putty from <https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html> from the site select software as per your system config depending on either 32-bit or 64-bit.



After downloading the .exe file for putty directly open the application and give host name as below



The click OK. So once in your in it will prompt for user name (Login as : ) and Password. You need to provide the WRDS credentials here. I don’t recommend using putty for you as you already have PC SAS installed so all the programs needs to be run from local PC SAS instead of putty. The reason to install putty is to check if you can open any TAQ database files using VIM editor so that it will determine if you access or not. For example if you want to know if you have access to a file named **nbbom\_20121231.sas7bdat**. In putty you need to navigate to location **/wrdslin/nyse/sasdata/taqms/nbbo** using command as below

**cd /wrdslin/nyse/sasdata/taqms/nbbo** then after thattype command

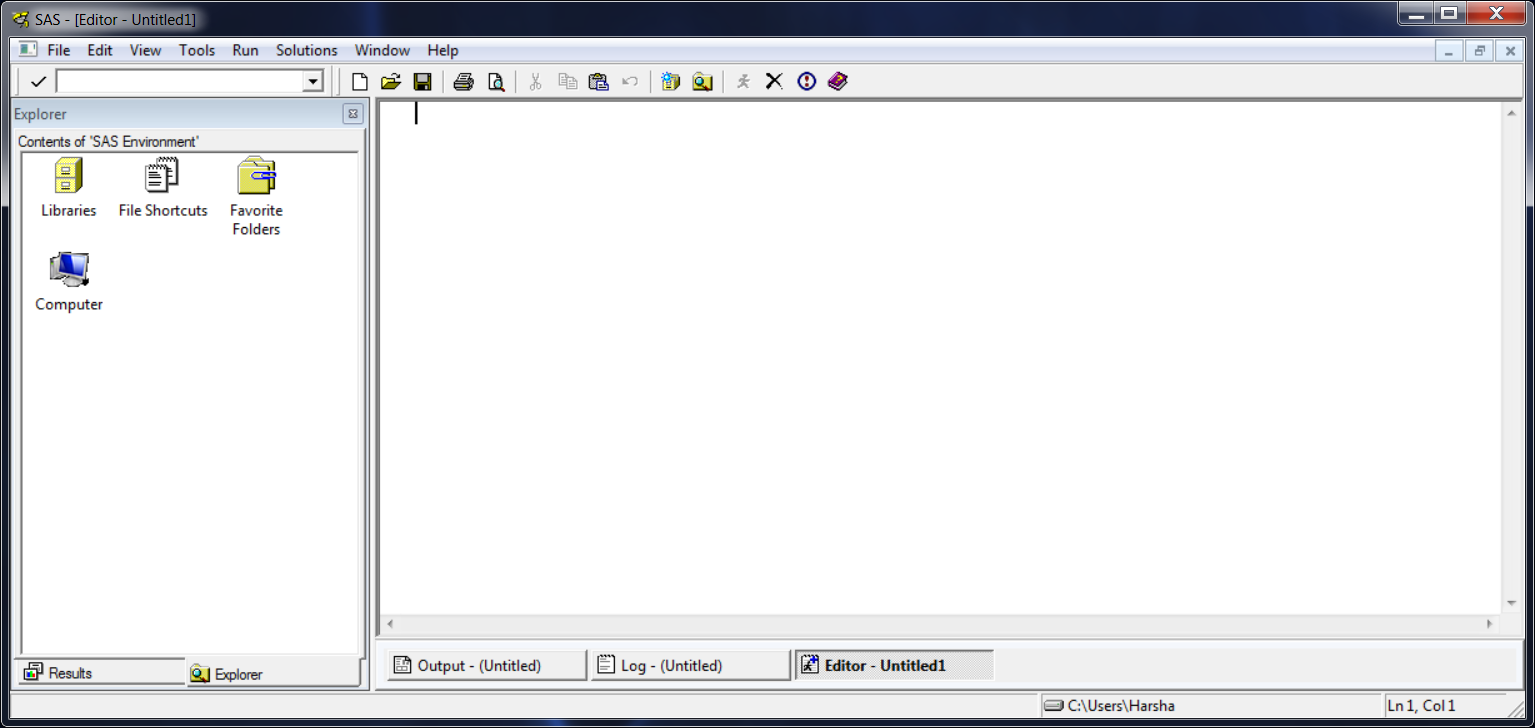
**vim nbbom\_20121231.sas7bdat** so if you don’t have access it will show as permission denied at the bottom of the screen as shown below.



If you have access to that file it will open that file and show the content. If you want to come out of VIM editor you need press **ESC** key and type **:q!** and press **ENTER** key. To navigate to other directories in Putty you need to have knowledge on Linux commands.

* **PC SAS**

**PC SAS is used to run the SAS programs. I am not explaining how to install the PC SAS as I assume PC SAS is already present in your machine and I will directly jump to how we can use it. When you open SAS software you will find three windows such as Editor, Log and Output.**



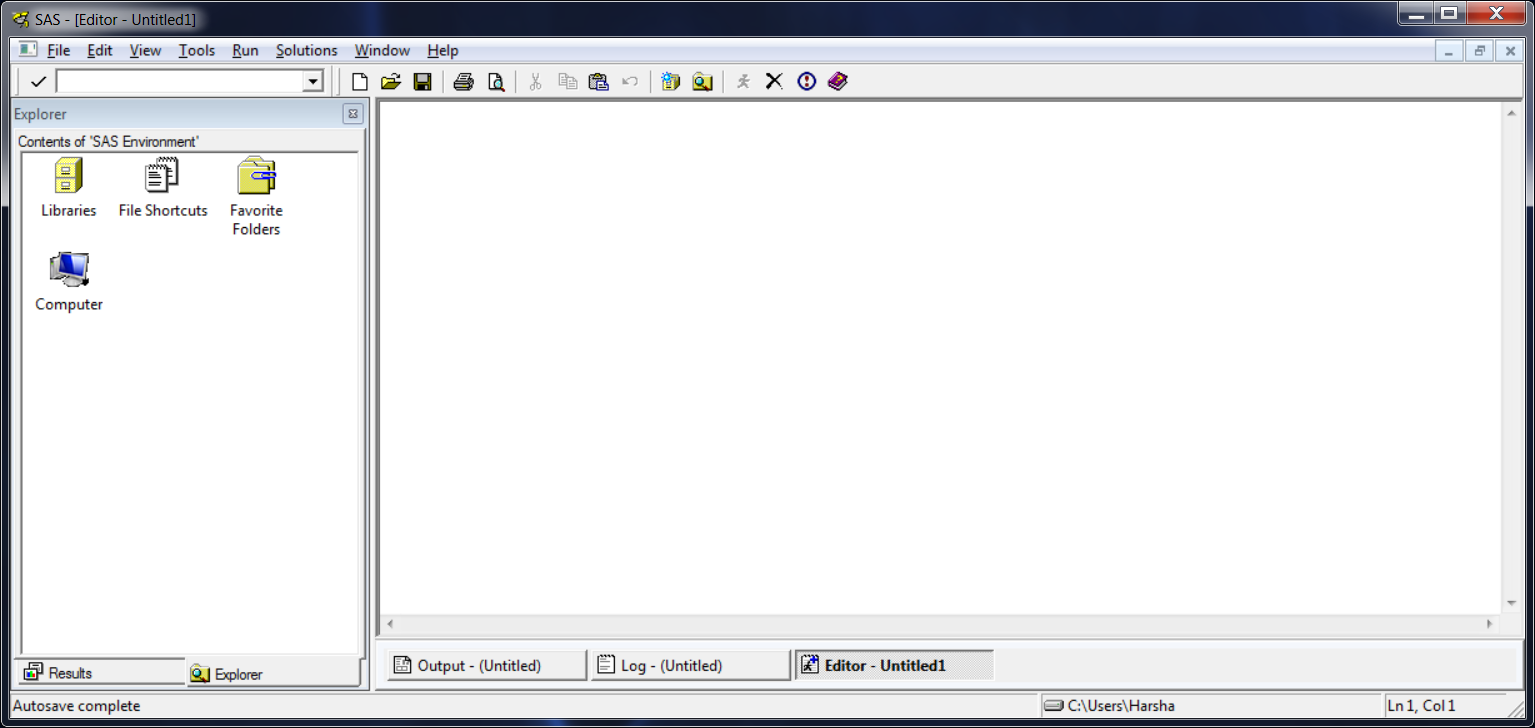
Editor is a window where you can write SAS programs or already existing SAS programs when opened from File menu will be shown in the editor window and whenever a program is run it creates a log and output. All the code statistics like how many records are processed and which input files and output files are used, ERRORS/Warnings are shown in the log window. If there is any print procedure executed in the SAS program it will print to output window by default.

**Step 2:**

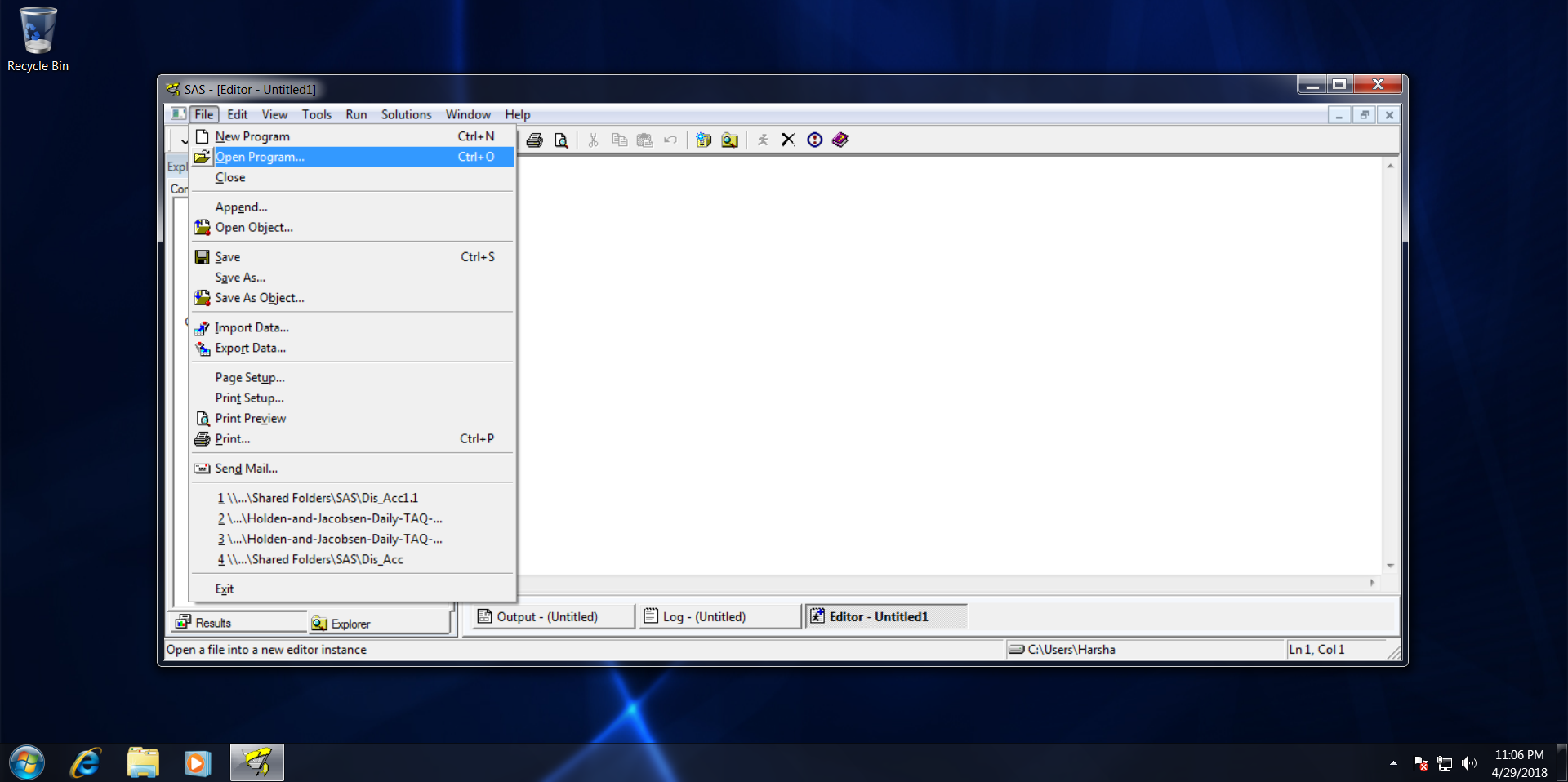
Moving to the next step I will explain how to run the codes from PC-SAS by connecting to WRDS TAQ database and execute the codes remotely.

**Executing Holden and Jacobsen Codes:**

Firstly, open the PC-SAS software in your local machine:



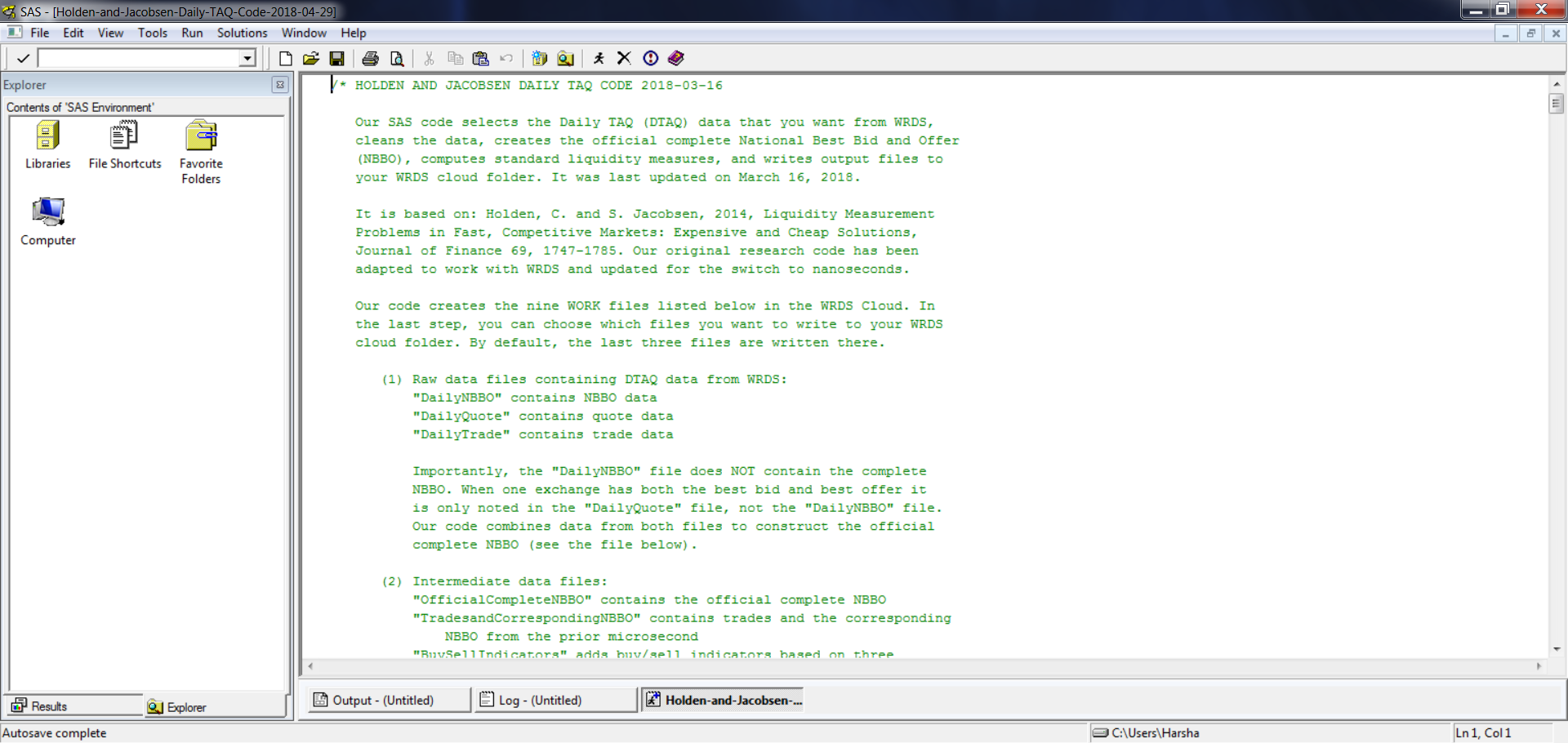
Then from File menu at top left area click open Open Program



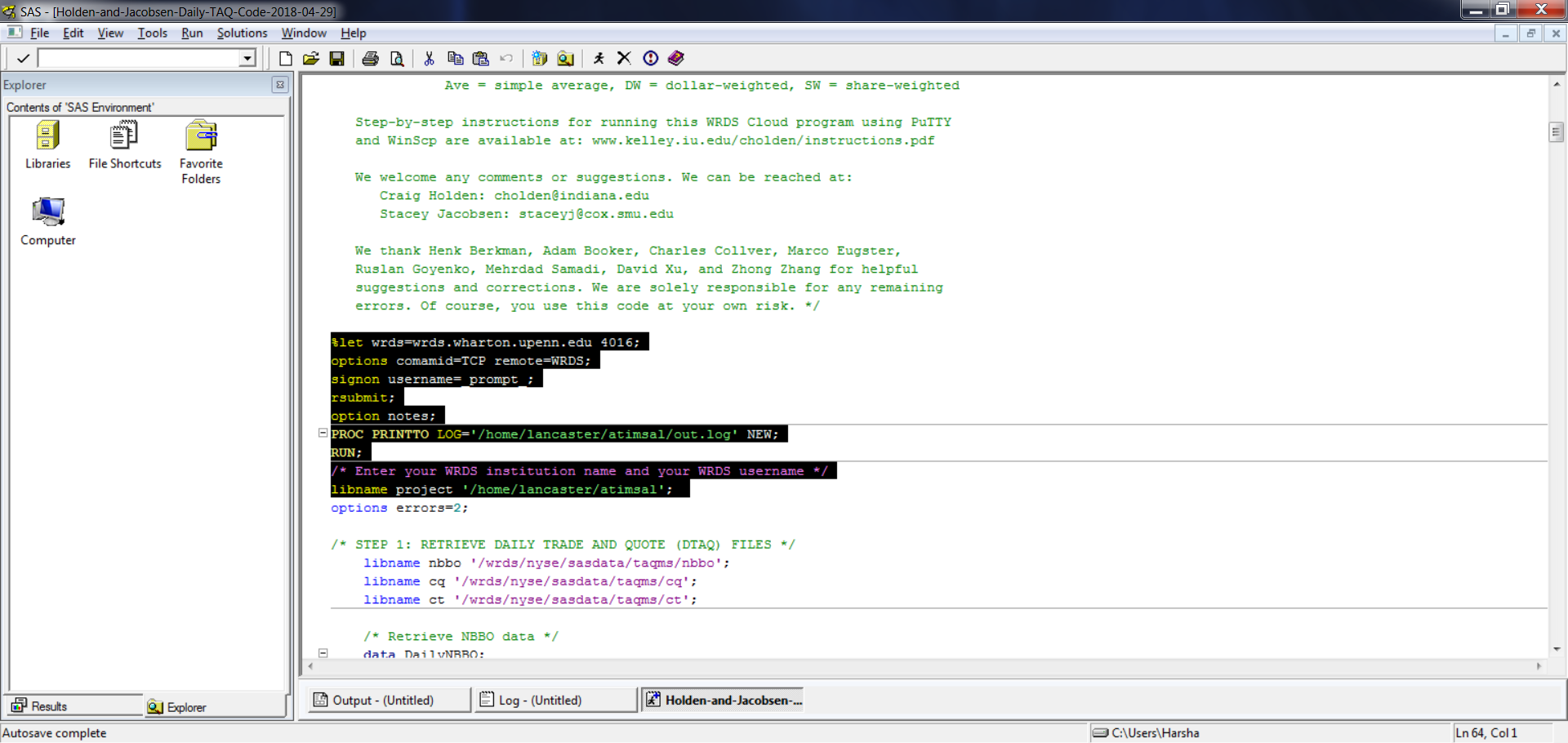
After clicking the Open Program navigate to the respective folder where you have saved your SAS program. The SAS programs are files with .sas extension. So here in this case we will open the file named

Holden-and-Jacobsen-Daily-TAQ-Code-2018-04-29.sas This program I have modified as per the requirements for TAQ database extraction by myself and will share with you with this document.

After opening the program it will look as below:



In the code at line 55 I have modified/added some codes of lines in the file as show below:



These are permanent changes to the code and shouldn’t be changed from your end. Lines 55-58 will instruct program to connect directly to WRDS site. Lines 60-61 shows that all the log of this SAS program execution will be written to a log present at location /home/lancaster/atimsal with name as Daily.log. So whenever you execute the program it will write the log to Daily.log instead of Log window on your machine. I have created this log as it will be easier to trace errors and debug the code in case of any issues. Another change which I have done is at line 63 libname project '/home/lancaster/atimsal'; where all the final output files are written to your home directory that is **/home/lancaster/atimsal** instead of being in SAS work directory. Just give you a clear idea where the TAQ datasets are present here are the locations which are indicated in the program.

libname nbbo '/wrds/nyse/sasdata/taqms/nbbo';

libname cq '/wrds/nyse/sasdata/taqms/cq';

libname ct '/wrds/nyse/sasdata/taqms/ct';

/wrds/nyse/sasdata/taqms/nbbo 🡺 Has all the NBBO files.

/wrds/nyse/sasdata/taqms/cq 🡺 Has all the Quote files.

/wrds/nyse/sasdata/taqms/ct 🡺 Has all the trade files.

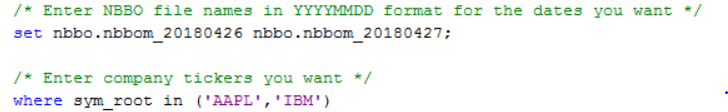
So you can navigate to those locations using WINSCP as I have already discussed in the WINSCP section installation and usage. We can navigate to that locations strongly recommend not to do as there are large number of files so it may hung-up the software.

Moving further in the program you can see at line 75 where we need to give NBBO input files. It can be either single or multiple files.

set nbbo.nbbom\_20180426 nbbo.nbbom\_20180427;

In the above case we are giving multiple input files for two different dates such as 20180426 and 20180427 in this way if you want to give another file you can give as below assuming 20180423 is a trading day and not a holiday.

set nbbo.nbbom\_20180426 nbbo.nbbom\_20180427 nbbo.nbbom\_20180423;



Next we have the SYMBOLS code which need to selected on line 78.

where sym\_root in ('AAPL','IBM')

If you want to add further symbols you need to add a comma and then symbol name in the single quotes within the parenthesis. Below I have shown how to add the symbol ‘AAT’.

where sym\_root in ('AAPL','IBM', 'AAT')

In the same way if you want to delete any unwanted symbols just remove them from parenthesis. Below I have shown the code if you want to do run only for ‘AAPL’ symbol.

where sym\_root in ('AAPL')

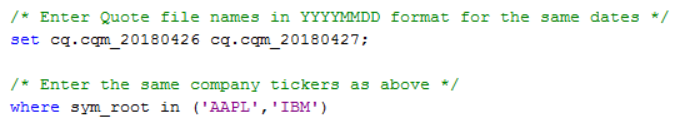
Do the same changes as we have done for NBBO in QUOTES and TRADES too.

Lines 101 and 104 are for quotes.

set cq.cqm\_20180426 cq.cqm\_20180427;

/\* Enter the same company tickers as above \*/

where sym\_root in ('AAPL','IBM')

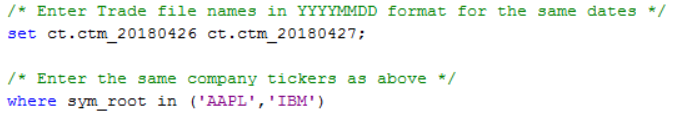


and lines 127 and 131 are for trades.

set ct.ctm\_20180426 ct.ctm\_20180427;

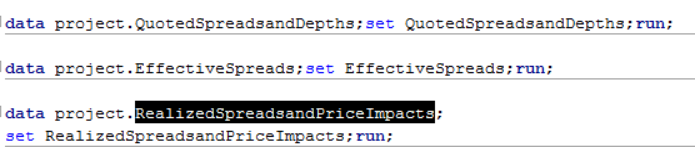
/\* Enter the same company tickers as above \*/

where sym\_root in ('AAPL','IBM')



All the files and Symbols for NBBO , Quotes and Trades should match exactly. Means if we are giving input files for NBBO as files of 26th and 27th April, 2018 then we need to give same files for Quotes and Trades sections too. And if we are giving symbols as ‘AAPL’ and ‘IBM’ in NBBO section we need to give same symbols in Quotes and Trades too.

From line 766 – 771 will have the final files created to the home directory.



The final files are

**QuotedSpreadsandDepths**

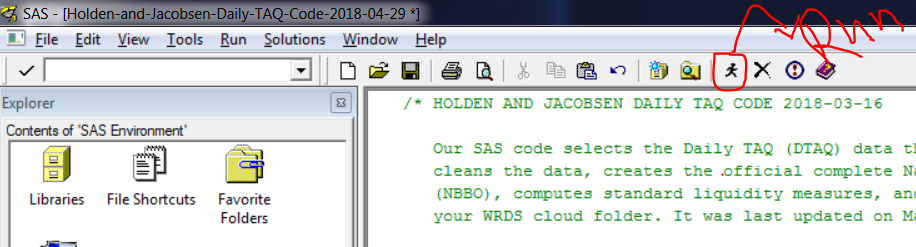
**EffectiveSpreads**

**RealizedSpreadsandPriceImpacts**

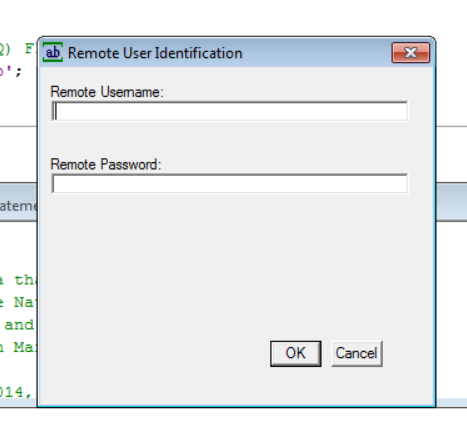
Lines 772 – 774 have been added to end the remote submission to WRDS site.

SAS Code Execution:

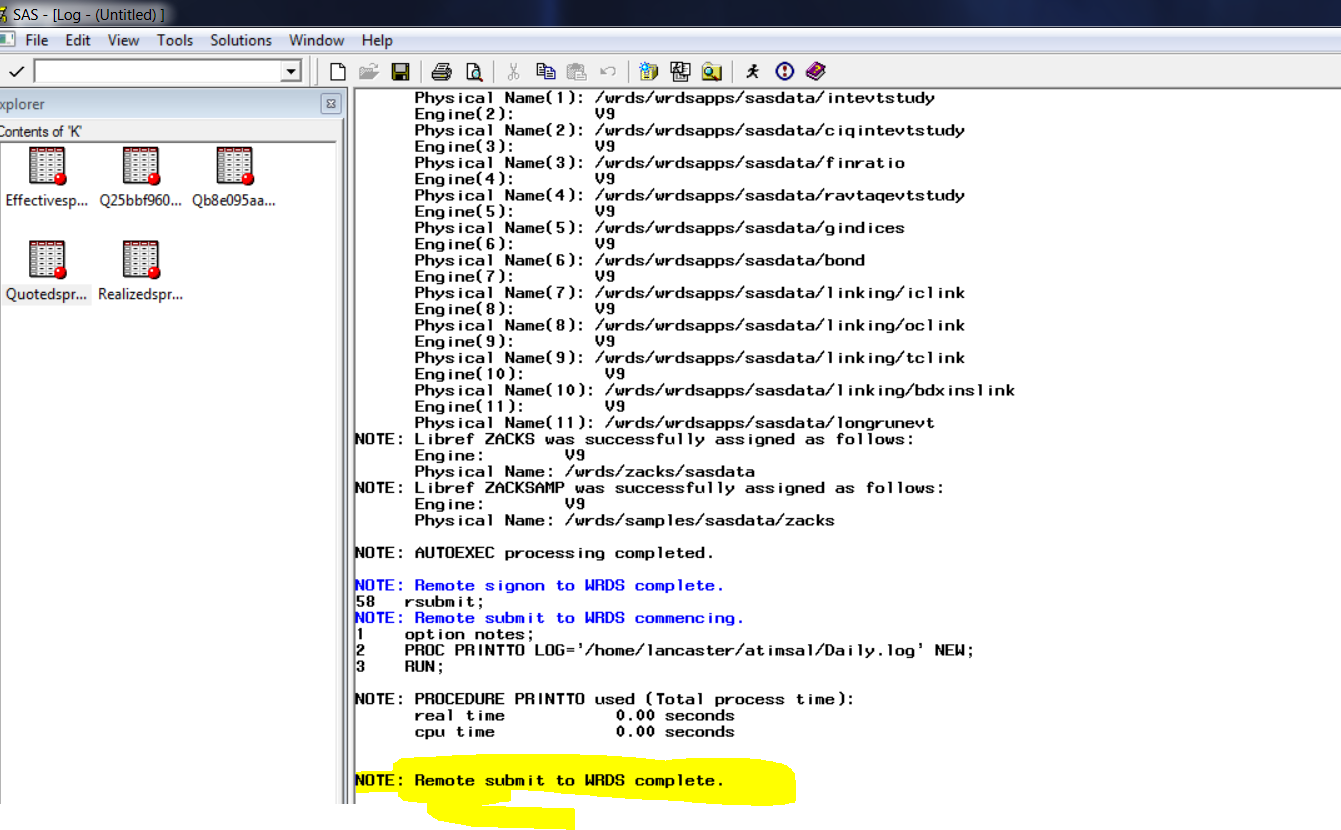
After editing the above SAS code we can run the code by clicking on the icon as indicated below in the red circle. Or Alternatively you can press F3 key.



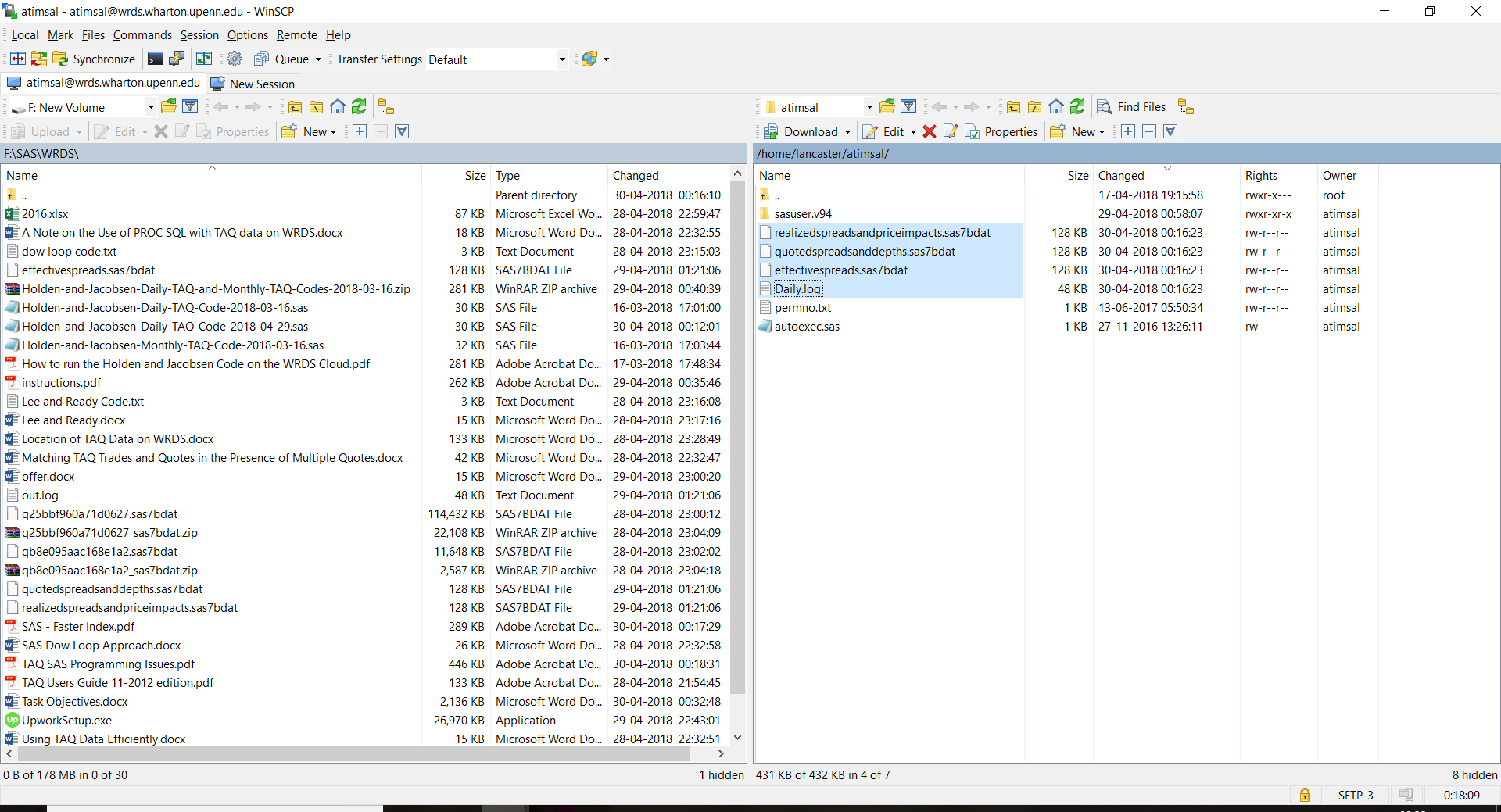
Once you click on Run button then you will prompted for your WRDS credentials like below



Here you need to enter your Username and Password and press OK button. After pressing OK button the actual processing on WRDS site starts and after successful completion of the execution it would create 3 SAS datasets in the home directory and a log named Daily.log. You can view your SAS Window log and check if the execution on remote site is completed or not as shown below



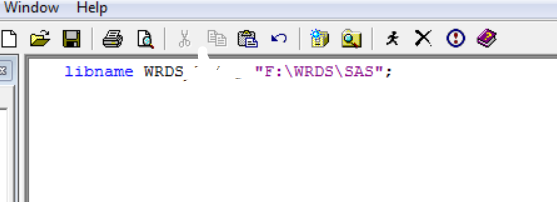
So as highlighted in yellow it says Remote Submit to WRDS is complete shows the execution is complete. After completing the execution open WINSCP and connect to WRDS site and in the home directory you can see three SAS datasets with extension as .sas7bdat and a SAS log file named Daily.log as shown below.



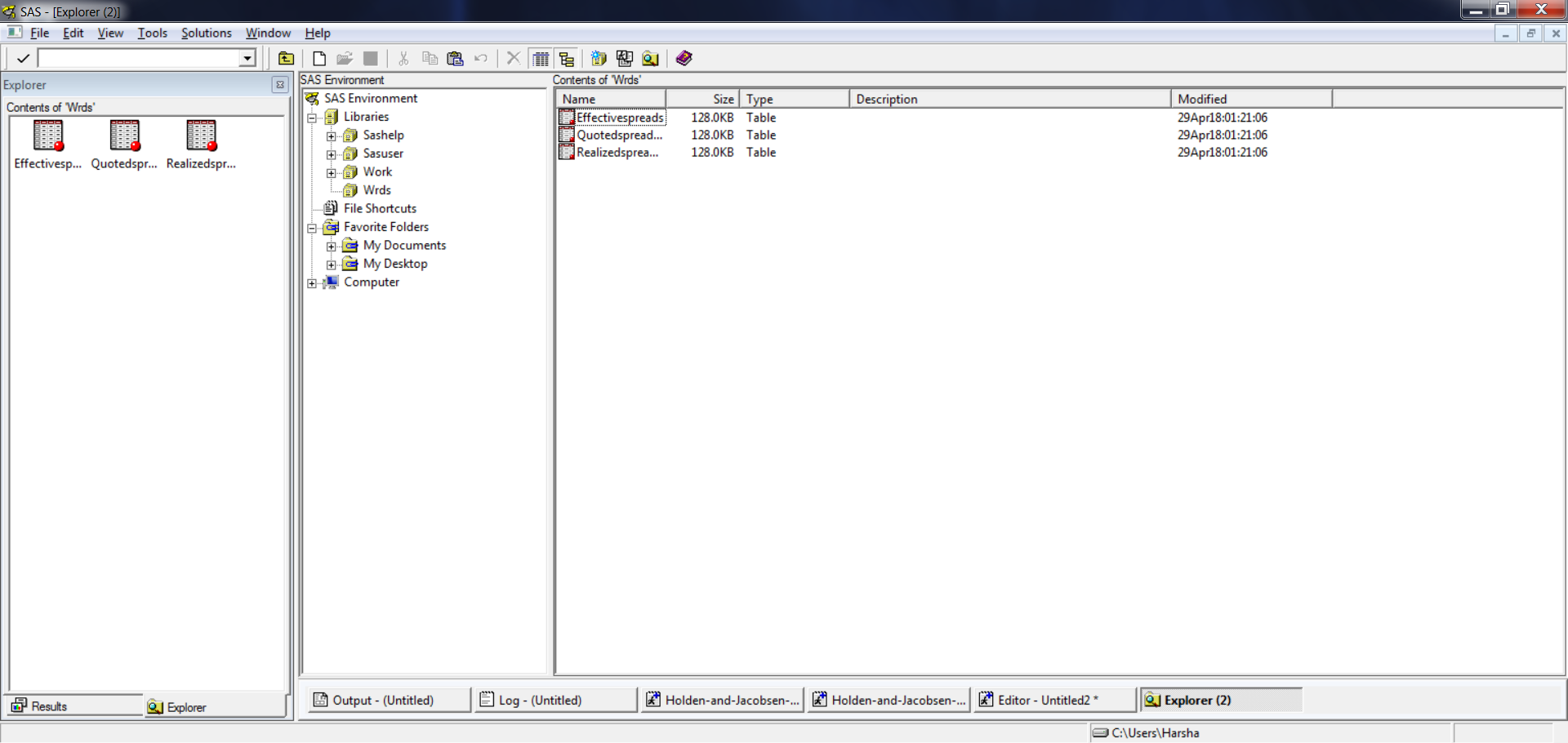
You can click the Daily.log file and open it and search if there are any errors. Press CTRL+F and find for ERROR: in the search box if there are any errors let me know I can help to fix. If you run the attached version of program there should be no errors in the log. And for viewing the SAS datasets you need to drag the .sas7bdat files from right side to left side of WINSCP . Before that setup a local directory in WINSCP. Here if you see above screen shot F:\WRDS\SAS is the local directory in my PC and once you drag files from right side to left window you can have final SAS outputs in your local machine(in my machine it is F:\WRDS\SAS).

After you are verified the log that there are no errors and moved the files from WRDS site to local machine directory now in order to view those files in SAS you need write below code in new SAS program window.(Open SAS Software, File Menu > New Program)

libname WRDS "F:\WRDS\SAS"; 🡺 Here instead of my local directory you need to give yours where you have dragged the files to through WINSCP.

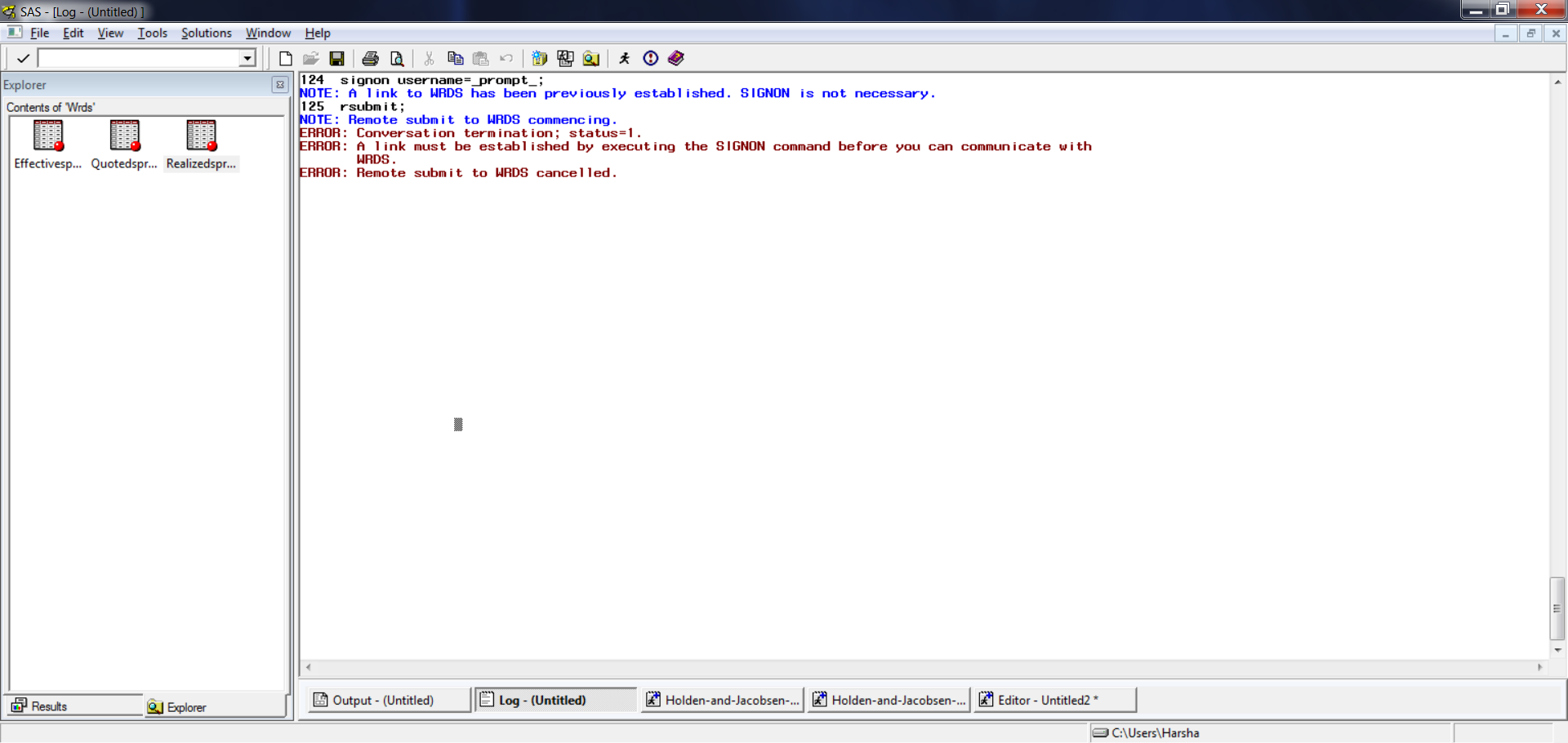


And execute the above code. It will create a new library in SAS which will have all the SAS datasets you moved from WRDS site through WINSCP. In order to view those files and open in SAS you need to open Explorer through View>Explorer and click on WRDS library you have created above it should show like below.



As you see above there are three datasets click on any of them to view the data. Ideally the datasets will have total number of symbols \* total number of days supplied in the program. For example here we have give 'AAPL','IBM' as symbols and 26th , 27th April files as input so it would have 2\*2= 4 records in each of three datasets.

At times you may get this error while executing the code in Log window.

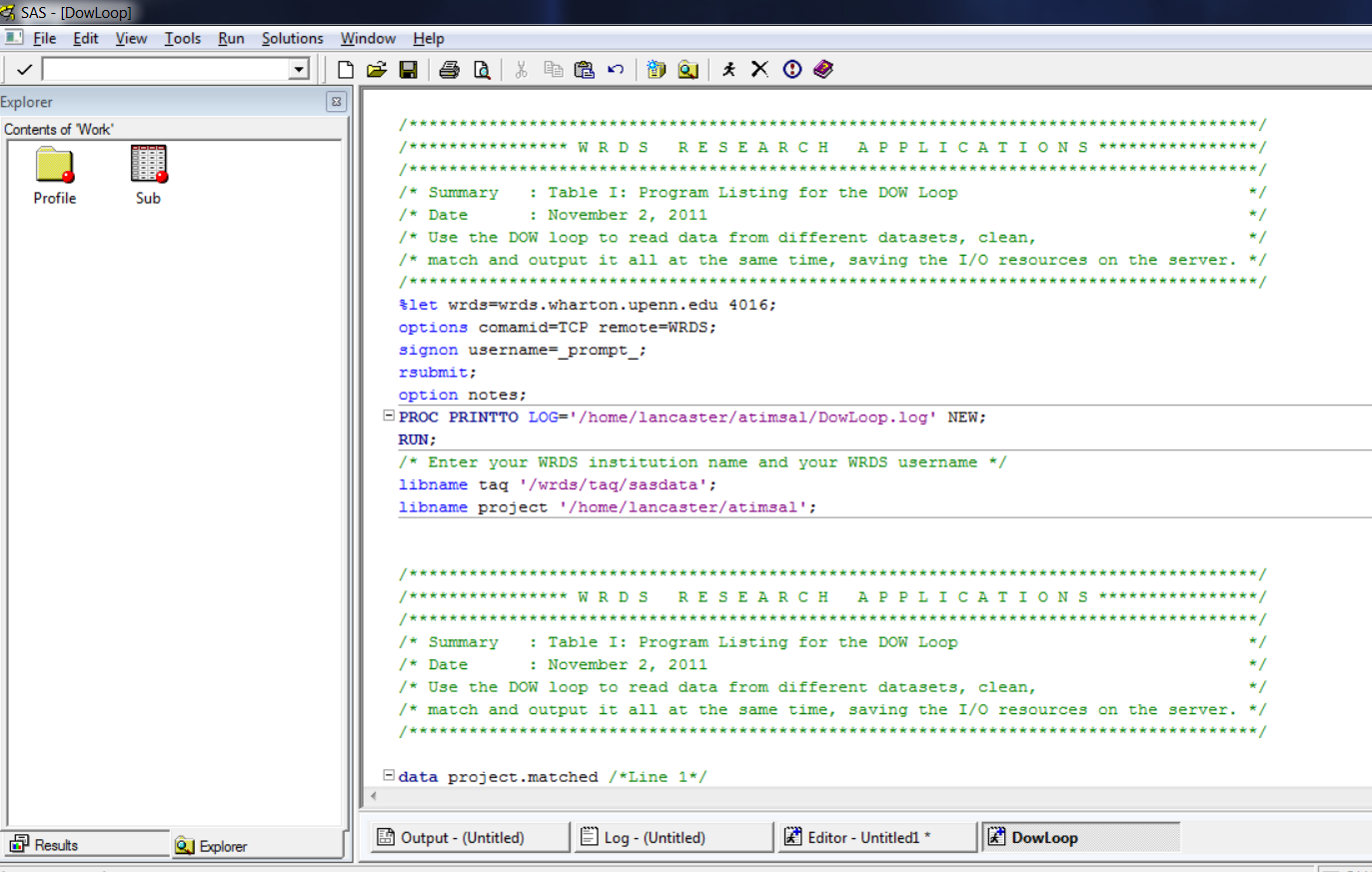


This indicates that either connection is broken or you need to rerun again by entering credentials.

Below are some of the program statistics which are generated by the code:

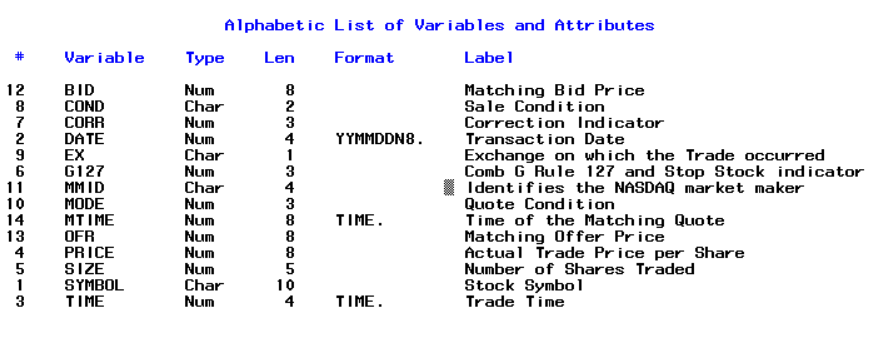
* Program has taken approximately 5 minutes to complete the execution for 2 days data and for 2 SYMBOLS.
* The program created 3 output datasets and each has 4 records that is SYMBOLS \* NUMBER OF DAYS of data.
* The program first extracts the data from NBBO, Quote and Trades datasets then combines NBBO with Quotes and then NBBO with Trades and then merges all three together to create aggregate fields.

**Dow Loop Code execution Steps:**

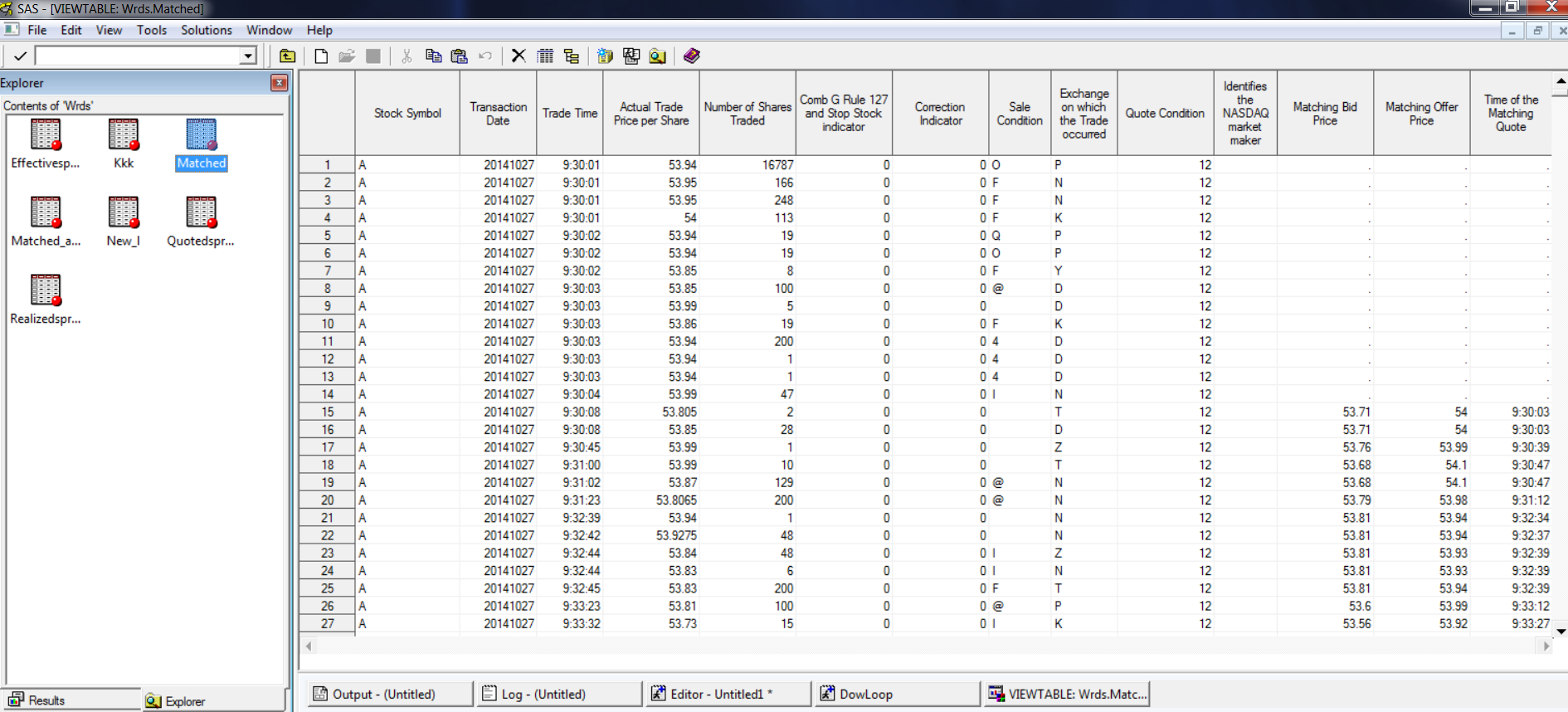
The DowLoop code can be executed by running DowLoop.sas code in PC-SAS. The program produces exactly one output file which is named as matched.sas7bdat in your home directory that is /home/Lancaster/atimsal and can be viewed by WINSCP as discussed in step1. You cannot open the SAS file in WINSCP and the file needs to be transferred to local machine by dragging from right pane that is WRDS site to left pane that is local machine. Below is screen shot of the code which is executed. The code match merges the data from Trade dataset and Quote dataset on keys such as Symbol, date and time of trades and produces a final output named as matched in the home directory. 

Below are the statistics which will help us to run the program in an optimized way without running into issues like Disk Usage/CPU, etc…

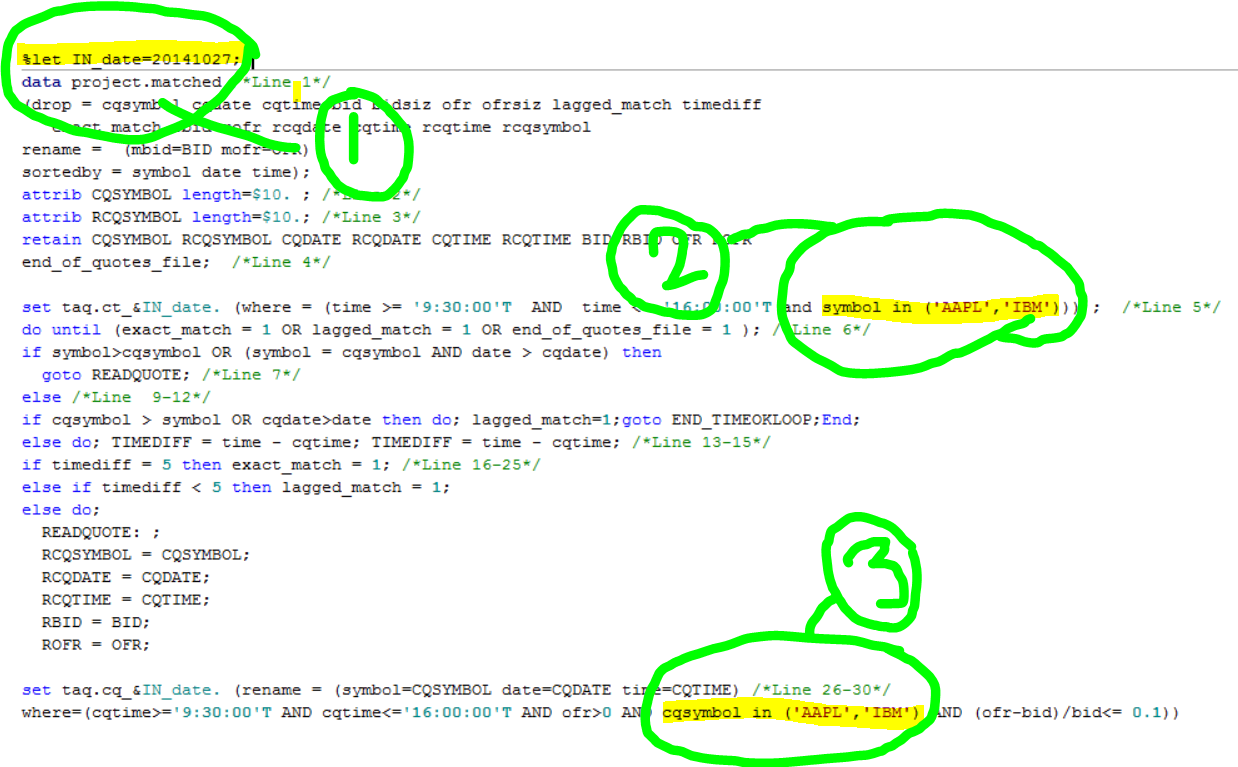
* Code when run for one day(20141027) that is without giving any specific symbol as input(means there is no filter on SYMBOL column from trades and quotes) it has taken approximately 7 minutes to execute and final matched dataset is created with 29M records and size is around 2GB.
* The final matched dataset had data for approximately 7800 SYMBOLS.
* Below are the columns created in the matched dataset.



* Below is the screenshot of the dataset.



* As we have already created a matched dataset of 2GB in the home directory whenever we rerun any other codes the system ran into memory issues so its always advisable to clear the older files when running the DowLoop code again for all symbols for one day.
* Next step as an experiment I have created a program to run only for one SYMBOL that is ‘AAPL’ as per program running stats it has take approximately 1.36 seconds for one symbol for one day and final matched dataset is around 12MB file with 175K records in it and columns are same as above case. Please note this on only for one day that is 20141027.
* Below are the changes that needs to be done in the code while running DowLoop code. The highlighted in yellow color are the places where we need to change the code:



In the first yellow colored portion indicates the date for which you have to run the code. So, if you want to run for today that is it should look like this **%let IN\_date=20180105;** instead of **%let IN\_date=20141027;**

The second and third highlighted areas we can mention symbols or which we need to run match merge for trades and quotes. If you want to run only for AAPL then the second area should be **symbol in (‘AAPL’)** and third area should be **cqsymbol in (‘AAPL’)** like wise you can provide as many number of symbols as you would like within parenthesis all symbols separated by commas. In order run for more symbols that is 3000 we need to change the code to accommodate a way to read the input file containing all symbols and then those symbols should be input in areas 2 and 3 as highlighted above. So the input file will only changed as per the need with number of symbols. In the same way we can also make the code to accommodate the dynamic dates in the area 1 as highlighted above.

* After you run the DowLoop code the log will be created in the home directory with name as DowLoop.log after each ad every run check if there are any errors. This can be done by opening DowLoop.log and doing CTRL+F and the search for **ERROR:**.