FINANCIAL STATISTICS
2021-10-29

Time:	14.00 - 19.30
Place:	*Data Hall 36, Albano Hus 4*
Approved aid:	All notes and course materials. Provide citations if you use out-side sources.
	You may copy code from your home assignment. However, you are not allowed to copy written explanations from your home assignment.

Questions about content of the exam: send e-mails to the following two e-mails
andriy.andreev@stat.su.se and ulf.hognas@stat.su.se

IMPORTANT: use "NewCode" for submission/uploading your answers

You are expected to upload R-packages and link the corresponding libraries:

> *install.packages("forecast")*
> *library(forecast)*
> *install.packages("tseries")*
> *library(tseries)*
> *install.packages("lmtest")*
> *library(lmtest)*
> *install.packages("car")*
> *library(car)*
> *install.packages("semTools")*
> *library(semTools)*
> *install.packages("DescTools")*
> *library(DescTools)*
>
> *install.packages("FinTS")*
> *FinTS::ArchTest()*

If some package you want to use is missing from the provided list, indicate clearly what package and for what purpose you use. No IT help will be provided for this step, except for if R-Studio does not work as expected.

If you ask a question during exam, you will get a clarifying answer about the content of the exam questions, not how to solve them. Allow for at least one hour to get your answer: the answers (if any) will be given at 15:30 and possibly at 16:00. No questions sent in after 15:30 will be answered.

Please, check your "registered" e-mails and the homepage on Athena at 15:00 for "updates" (if any).

In exceptional case that you cannot submit your work at the very end of the exam, contact tenta@stat.su.se. You are expected to practice submitting couple days before the exam: instructions have been given earlier.

- **Part I (Time Series)**

    You are asked to analyze **WEEKLY** stock data using R and submit a written assignment in essay form. Your essay should contain output from R as well as clear interpretation of results. You are also asked to attach R code for the outputs you use to motivate your answers. In case no R code is attached, the final grade will be reduced.

- **Part II (Regression)**

    You are asked to analyze the data using R and submit a written assignment in essay form. Your essay should contain output from R as well as clear interpretation of results. You are also asked to attach R code for the output you use to motivate your answers.

    Reminder: all the sheets you are submitting should contain your anonymous code and be ordered the way you want your answers are read.

    Upload Deadline: 19:30

- **Grading, minimum requirements:**
    - Pass, E or better: answer Part I a-d (Part II can compensate for lacking answer in d)
    - C or better: fully answer Part I a-d, good part of e and make a fair attempt on both f-g, and on Part II
    - A: answer Part I in full with minor deficiencies. Answer Part II in a satisfactory way

These are preliminary requirements. The examiner might adjust the scale upward, if necessary.

- NOTE! Fx and F are failing grades that require re-examination. Students who receive the grade Fx or F cannot supplement for a higher grade.
- If you are not satisfied with your grade, you may contact the examiner for further instructions.
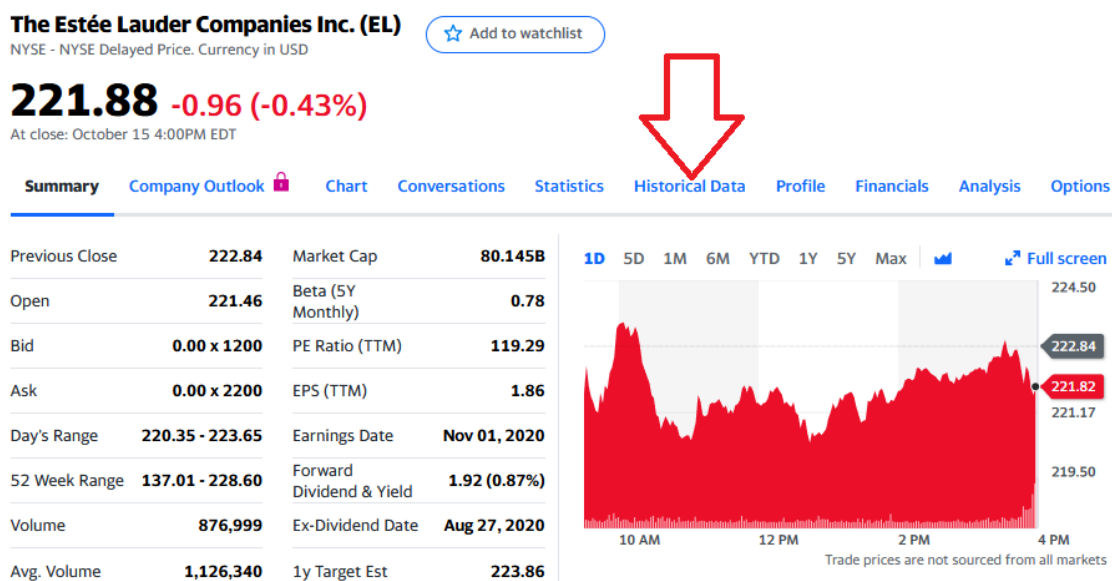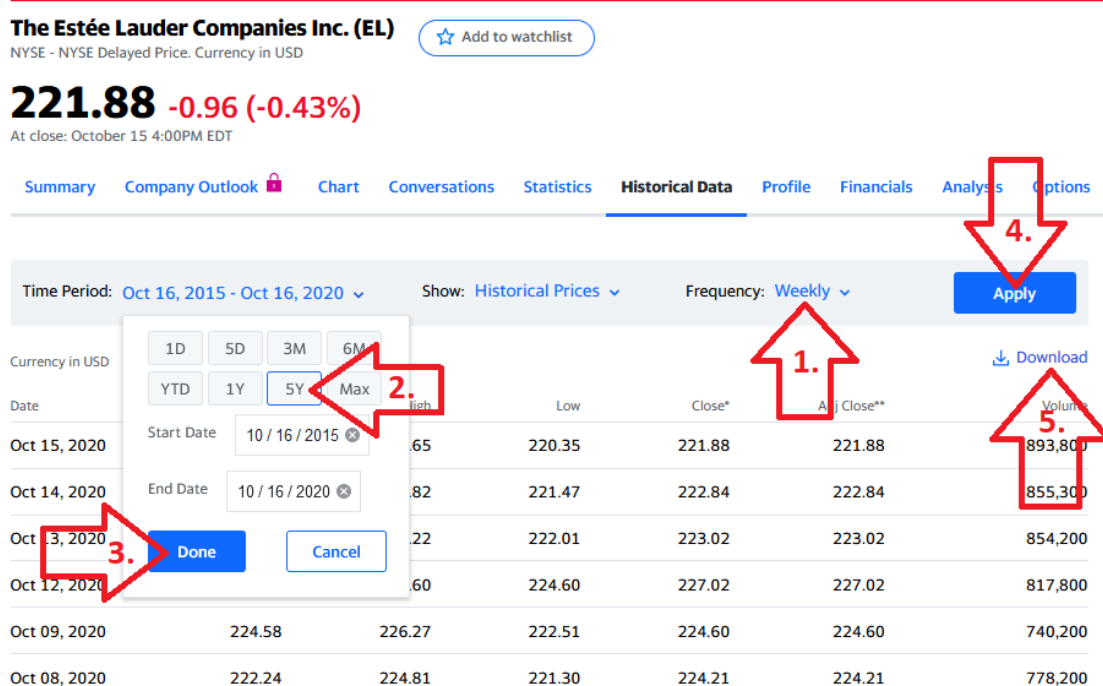
GOOD LUCK!

Part I, Data

Below, you will find a selection of 50 companies from the S&P 500. Your data corresponds to the number on a "NewCode" that is matched to your anonymous code. For instance, if you have the NewCode 003-XXX, your data is "BLL". Follow these instructions to download your stock:

1. Go to https://finance.yahoo.com/

2. Enter the "Symbol" for your stock into the Yahoo! Finance search field and press enter. *For example, if your stock is The Estee Lauder Companies Inc. enter "EL" into the search field and press enter.*
3. Click historical data:

4. Change Frequency to "**Weekly**"; Click Time Period, choose "**5Y**", and click "**Done**"; Click "**Apply**"; finally, click "**Download**." Make sure that you have five years of weekly data, or slightly more (263 observations or so). If your stock does not have 5 years of data, choose "**Max.**" In this case, you will of course have less than 263 observations.

5. You may read the data into R any way you want; the following step is just a suggestion. Open a new .R-file and paste in the following code:

```
EL <- read.csv("EL.csv")
```

This assumes that you have copied the downloaded stock data file to your work directory. Remember that you can find out what your work directory is with the command

getwd()

You can list the files in your work directory with

list.files()

If you are having trouble loading the file, first make sure that your file is in the work directory. Save the .R-file and run it. Then verify that the data is correctly read into R.

Use Adj.Close price.

NewCode	Code	Symbol	Name
001-UBG	311-UBG-KHB	SLG	SL Green Realty
002-GZR	311-GZR-EFF	NDAQ	Nasdaq Inc.
003-ZLX	311-ZLX-ERP	BLL	Ball Corp
004-MPO	311-MPO-YDF	BKNG	Booking Holdings Inc
005-XFS	311-XFS-LGJ	PKI	PerkinElmer
006-FYT	311-FYT-OHX	DISCA	Discovery Inc. (Series A)
007-XKC	311-XKC-MUY	MS	Morgan Stanley
008-YPH	311-YPH-KMA	FIS	Fidelity National Information Services
009-DRG	311-DRG-HAS	AMD	Advanced Micro Devices
010-GZX	311-GZX-LBZ	DOW	Dow Inc.
011-GZA	311-GZA-FXP	SLG	SL Green Realty
012-ELK	311-ELK-JZK	NDAQ	Nasdaq Inc.
013-OLF	311-OLF-JFW	BLL	Ball Corp
014-BSH	311-BSH-ASK	BKNG	Booking Holdings Inc
015-XSU	311-XSU-SYA	PKI	PerkinElmer
016-RPY	311-RPY-EFC	DISCA	Discovery Inc. (Series A)
017-DPW	311-DPW-XEE	MS	Morgan Stanley
018-NDD	311-NDD-XNN	FIS	Fidelity National Information Services
019-URT	311-URT-YKT	AMD	Advanced Micro Devices
020-YJN	311-YJN-MYP	DOW	Dow Inc.
021-DKS	311-DKS-CSJ	SLG	SL Green Realty
022-FNY	311-FNY-GKU	NDAQ	Nasdaq Inc.
023-JFZ	311-JFZ-AJP	BLL	Ball Corp
024-HFK	311-HFK-ZER	BKNG	Booking Holdings Inc
025-WKJ	311-WKJ-WRB	PKI	PerkinElmer
026-XMP	311-XMP-SLG	DISCA	Discovery Inc. (Series A)
027-AZG	311-AZG-YMM	MS	Morgan Stanley
028-JDS	311-JDS-ARS	FIS	Fidelity National Information Services
029-GKG	311-GKG-NEL	AMD	Advanced Micro Devices
030-WGL	311-WGL-KFR	DOW	Dow Inc.
031-EUB	311-EUB-JOT	SLG	SL Green Realty
032-SHK	311-SHK-BGX	NDAQ	Nasdaq Inc.
033-WHY	311-WHY-UTO	BLL	Ball Corp

034-BNX	311-BNX-JLM	BKNG	Booking Holdings Inc
035-ULZ	311-ULZ-KDO	PKI	PerkinElmer
036-UYX	311-UYX-WTO	DISCA	Discovery Inc. (Series A)
037-OCT	311-OCT-CNU	MS	Morgan Stanley
038-EBJ	311-EBJ-ZRP	FIS	Fidelity National Information Services
039		AMD	Advanced Micro Devices
040		DOW	Dow Inc.
041		SLG	SL Green Realty
042		NDAQ	Nasdaq Inc.
043		BLL	Ball Corp
044		BKNG	Booking Holdings Inc
045		PKI	PerkinElmer
046		DISCA	Discovery Inc. (Series A)
047		MS	Morgan Stanley
048		FIS	Fidelity National Information Services
049		AMD	Advanced Micro Devices
050		DOW	Dow Inc.

Part I, Instructions for the Time Series Part

The goal of the assignment in Part I is to perform the analysis of your assigned stock in R and then write a short essay answering the questions below. Make one headline in the essay for each part that you answer, e.g. "a. Introduction and Trend" and "b. Stationarity." It is important that you include plots in your assignment.

(a) Describe the time series with an appropriate diagram. Comment on the diagram. Is there a visible trend? You do not need to consider seasonal effects this time.

   For steps b, c, and d, leave out the last 3 observations. The last four observations will be referred to as "the testing set" here.

(b) Stationarity. Use a formal test for stationarity. If it is clear from the diagram that your data is non-stationary, you may apply an appropriate transformation without testing first. State the hypotheses and interpret the outcome of the test. Use 5% significance level. If the data does not seem to be stationary even after transformation, attempt another transformation and repeat this step.

(c) Plot the ACF and PACF plots. Explain what the plots show. For example, explain what the value of the first bar in the ACF plot mean. Use these plots as basis for your choices in (d):

(d) Choose at least four ARIMA models to analyze the data (leaving out the testing set). One of these should be an ARIMA(0,d,0): indicate what "d" you have chosen and why. Provide output and AIC scores for each model. Interpret the results. Which model seems most promising? Should any of your models be disregarded completely?

(e) Choose the two "best" models from (d). Use the last 3 observations, i.e. your testing set, to calculate RMSE. Compare the two and interpret the difference.

(f) Pick "the best" model from question (e). Test that model for GARCH effect. State the hypothesis and explain the outcome of your test. Explain what the presence of GARCH effect would mean for your results in (d), even if you did not find any. Discuss briefly why estimating variance is important in financial forecasting.

(g) Pick "the best" model from question (e). Provide residual analysis of the model and report your conclusions.
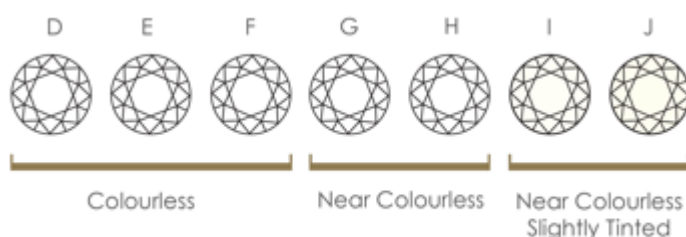

Part II (Regression Part)

The file **diamondsample.csv** contains data from a random sample of **200** round cut diamonds. The data set contains the following variables:

price: Estimated price in USD

carat: weight in carats (1 carat = 200 milligram)

Note: the charts have been cropped so that they only included the relevant color and clarity categories.
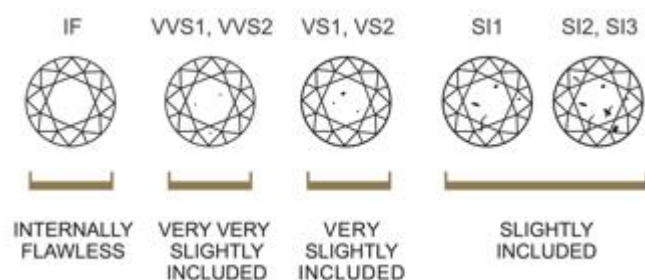
COLOR CATEGORIES

colorDEF: a dummy variable: "1" = the diamond belongs to color category D,E or F (the most desirable color categories)

colorGH: a dummy variable: "1" = the diamond belongs to color category G or H (the color categories below D,E, and F)

The base category for color is all categories less desirable than D, E, F, G, and H.

CLARITY CATEGORIES

clarityIF: a dummy variable: "1" = "internally flawless" - the best clarity category

clarityVVS: a dummy variable: "1" = "very very slightly included" - the second best clarity category

clarityVS: a dummy variable: "1" means "very slightly included" - the third best clarity category

The base category for color is all clarity rating less desirable than the above three ("slightly included")

Read the data set into R. First copy the data to your work directory (not to a folder inside your work directory; the file needs to be directly in your work directory). Since this is .csv format, you can use read.csv:

diamonds <- read.csv('diamondsample.csv')

a) Summarize the data using **summary(diamonds)**. What is the price range of the diamonds in the sample: the highest and the lowest price? Any other descriptive statistics you find important to report?

b) Create a scatter plot with **price** on the y-axis and **carat** on the x-axis. Do you believe that this relationship is linear? Explain your motivation.

   For parts (c), (d), and (e), you should use **summary(modelname)**, for example **summary(model1)**.

c) **Model 1** Estimate a regression model with the **price** as <u>dependent</u> variable and **carat**, **colorDEF**, **colorGH**, **clarityIF, clarityVVS,** and **clarityVS** as independent variables. Save the model under the name **model1**. Plot the residuals of Model 1 against the fitted y-values. Comment.

d) Create a new variable that is the natural log of price. Estimate a new model that we will call **Model 2.** Model 2 should have the natural log of the price as dependent variable. The independent variables should be the same as those in Model 1. Plot the residuals of the Model 2 against the fitted y-values. Comment.

e) Create new variables:
   carat2   this should be carat squared
   carat3   this should be carat cubed
   carat4   this should be carat cubed

   Add all or some of these as independent variables to the model in (d). Try adding

   i.   carat2 only

   ii. carat2 and carat 3

   iii. carat2, carat3, and carat4

   Use **summary()** and the p-values to choose the best model. Save this best model under some name. Plot the residuals of this model against the fitted y-values. Be sure to comment

on the residuals analysis of your best model. Did the polynomial terms improve the residuals?

f) Use model 3 to find the estimated price of a diamond with the following characteristics: 0.75 carat, color "D", and clarity "IF."