Outline of and Rubric for ECON370 Final Project

Drew Van Kuiken

University of North Carolina | ECON 370

Final Project

Overview

The goal of your final project is to complete an independent[1] data science project. This will consist of:

- A research question
- Finding some data (likely multiple datasets)
- Running some regressions
- Creating some graphs and descriptive tables
- A short (~5 slide) in-class presentation[2]
- A longer (~20 slide) presentation[3] which the grader and I will review, but which will not be presented live
- One or more R programs which provide the **full** workflow needed to replicate your results

I describe each component in detail below. As a general rule: if you have questions, please ask them. If you feel lost, ***please come to my office hours***. I am more than happy to be a resource for you and help you through problems as they arise.

Note: LLMs are ***allowed*** on your final project and PS3. Please include information (a slide in your presentation/a comment in your HW submission) about what you used the LLM to do!

Rubric

Category	Points
Conceptual	
Research Question	15
Has Data	15
Reasonable regression(s)	15
Relevant graphs and tables	15
Coding	
Code does what it's supposed to	15
Code runs without errors	15
Exhibits good programming practices	15
Appropriately commented	5
Sanely organized	5
In-class presentation	
Clear, cogent discussion of project	5
Introduction/Research Question slide	5
Data slide	5
Findings (1)	5
Findings (2)	5
What did we learn / why is this important?	5
Final presentation	
At least 15 slides	5
Logical flow through topics	10
Motivates project - why should we care?	15

Category	Points
Discusses drawbacks to project	15
At least 1 table and 1 graph	15
At least 1 regression	15
Subtotal	220

Note: Think of your in-class presentation as a discussion of "here's what I did, here's why it matters." Your extended slide deck should additionally answer "here's why I did what I did, here are some drawbacks and limitations."

Research Questions

Developing good research questions is an art and probably beyond the scope of this course as it stands. As a result, don't sweat your research question too much here. Mostly, I want you to identify an outcome of interest and one or more things to test against that outcome of interest. Most important is that you can identify a single question that forms the basis of your project. Note that you can make your project about whatever you're interested in. Don't feel like your project has to be about something formal and business-related.

Broadly, there are two types of questions you can ask: descriptive questions and causal questions.

Descriptive Questions

Descriptive questions answer how things are. They *describe* the world. A simple example: what is 2+2? A more advanced, personal, example: how do users of a major airline's website differ in their search and purchasing behavior based on observable characteristics?[4]

Descriptive questions are famously underappreciated in academia, but they're super important in the real world. You're welcome to pursue a descriptive project in this class. As a general rule of thumb, descriptive projects should focus more on the data side of things. Think about what kinds of datasets might be interesting to combine and what insights you might get from linking two topics together. In the case of our airline paper, the paper exists basically because the data is so cool and rare. It's not often you get to see behind the curtain of an airline's website (or really any major corporation's website).

Conversely, descriptive projects will frequently place less weight on the regression side of things. For the purposes of our final project, you should still have at least one regression (note that our airline paper has a couple regressions).

Think of your regression as a descriptive object.[5]

Causal Questions

Causal questions answer whether a relationship exists between two objects. They are *why* questions. What causes something to increase (or decrease)? A simple example: how does your college determine your mid-career earnings? More advanced, more personal: how does competition affect the technology that solar panel producers adopt and develop?[6]

There are many, many things to learn about causal inference that we have not covered. Josh Angrist, Guido Imbens, and David Card won the economics Nobel prize in 2021 for what's been termed the "causal revolution" in economics. They developed a huge suite of controversial tools for trying to manipulate observational data such that it looks and acts like experimental data. It's cool stuff! I encourage you to learn more about it in future studies.

For our purposes, final projects that answer causal questions will lean more into discussions of (a) your attempts to match on observables, and (b) how bias might be affecting your results. It's a little

less important that your data itself is interesting here. Instead, I'd like you to focus on asking exciting questions.

Finding Some Data

Finding data stinks. It's really hard and sometimes it costs like $20,000. Below, I've tried to list every public data source I can think of. If nothing here appeals to you, come to my office hours and we can try to figure out how to find data on something that does interest you!

Data Repositories:

- Data.gov[7]
- FRED[8]
- UNC Libraries[9]
- Kaggle[10]
  - Strict requirement: talk with me if you want to use Kaggle data.
- Bureau of Labor Statistics Employment, Productivity, Inflation, and Wages data
- ICPSR huge repository, mostly american data

- Google a product someone built and forgot about at Google, I assume
- IPUMS really cool surveys about demographics, prices, time use, higher ed, and health stuff

Topic-specific resources:

- Industrial Organization (I currently use some of these things)
  - Airplanes 10% sample of all domestic air travel passengers! Many other resources on the website too. Check out the T100 database as well.
  - Solar Panels Crazy survey of most solar panel owners in the domestic US
  - Health stuff Several databases maintained by UNC. Might not have enough time to

access these resources.

- Medical Expenditures Household and insurance summary data about healthcare spending over time

- Labor and other applied microeconomics
  - NLSY yearly survey of youth jobs/time use/crime/earnings/school stuff. Incredible resource

- Macroeconomics
  - I would probably stick to FRED? And additionally stuff from the Bureau of Labor Statistics

- Sports
  - Football: nflfastR
  - Basketball: hoopR - men's

college and professional basketball data, wehoop - women's college and professional basketball data

- Hockey: nhlapi
- Baseball: baseballr

- Something you scraped
- A Monte Carlo simulation you ran
  - Chat with me if you're interested in this, we can try to figure something out

Data Processing and Analysis

It's hard for me to be prescriptive here - what's required of you will depend on your question and your data. In general, it is hard for me to think of a project that wouldn't benefit from multiple data sources or datasets.[11] Additionally, your project will almost certainly involve "cleaning" the data. To do so, you should deal with missing or incorrect information, create variables where

needed, and normalize units where relevant.[12] Be thoughtful about how each observation in your data relates to the question you're asking.

Let's look at an example. Consider a project about solar panel manufacturers and technology choices. A bit of background: for our purposes, there are two main technologies used in solar panels, multicrystalline silicon and monocrystalline silicon. Both have been produced for the past 15 years, though monocrystalline silicon is becoming increasingly popular. Multicrystalline silicon is cheaper but less efficient. If consumers have preferences over price and efficiency, you can imagine producers wanting to sell one cheap, low-efficiency panel, and one expensive, high-efficiency panel.

Before thinking about a research question and formal regression model, we might want to start with some graphs. Over the past 15 years, for each manufacturer in the data, what did their newly developed products look like? How did the efficiency of their low-cost multicrystalline products and their high-cost monocrystalline products develop over time?

Let's take a look at the Tracking the Sun database to get a sense of this:

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────────
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyve
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>)
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
##
## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday,
##     yday, year
##
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
##
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
library(ggrepel)
```

```
### goal: create single-dimension plots of efficiencies by techno
###       firm identity for each year of the data. subset to firm
###       market share.
```

```
###
# note to students: the tracking the sun data file is saved to ou
# this code on your own machine. be careful though, the file is p
path <- '~/Desktop/Research/solar/data/raw/Tracking the Sun 2023
# Read in TTS data
tts_raw = read_csv(paste0(path,'/TTS_LBNL_public_file_27-Oct-2023
```

```
## Rows: 2990939 Columns: 81
## ── Column specification ────────────────────────────────────────
## Delimiter: ","
## chr (31): data_provider_1, data_provider_2, system_ID_1, syste
## dbl (50): PV_system_size_DC, total_installed_price, rebate_or_
##
## ℹ Use `spec()` to retrieve the full column specification for t
## ℹ Specify the column types or set `show_col_types = FALSE` to
```

```
glimpse(tts_raw)
```

```
## Rows: 2,990,939
## Columns: 81
## $ data_provider_1            <chr> "Arizona Public Service",
## $ data_provider_2            <chr> "-1", "-1", "-1", "-1", "-
## $ system_ID_1                <chr> "1", "2", "3", "4", "5", "
## $ system_ID_2                <chr> "-1", "-1", "-1", "-1", "-
## $ installation_date          <chr> "22-Dec-1995", "24-Jan-200
## $ PV_system_size_DC          <dbl> -1.000, 12.025, -1.000, 8.
## $ total_installed_price      <dbl> -1, -1, -1, -1, -1, -1, -1
## $ rebate_or_grant            <dbl> -1, -1, -1, -1, -1, -1, -1
## $ customer_segment           <chr> "RES", "RES", "RES", "RES"
## $ expansion_system           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ multiple_phase_system      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ TTS_link_ID                <chr> "-1", "-1", "-1", "-1", "-
## $ new_construction           <dbl> -1, -1, -1, -1, -1, -1, -1
## $ tracking                   <dbl> -1, -1, -1, -1, -1, -1, -1
## $ ground_mounted             <dbl> -1, -1, -1, -1, -1, -1, -1
## $ zip_code                   <chr> "85262", "85338", "85320",
## $ city                       <chr> "Scottsdale", "Goodyear",
## $ state                      <chr> "AZ", "AZ", "AZ", "AZ", "A
## $ utility_service_territory  <chr> "Arizona Public Service",
## $ third_party_owned          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,
```

```
## $ installer_name              <chr> "-1", "Unknown", "-1", "Ti
## $ self_installed              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ azimuth_1                   <dbl> -1, -1, -1, -1, -1, -1, -1
## $ azimuth_2                   <dbl> -1, -1, -1, -1, -1, -1, -1
## $ azimuth_3                   <dbl> -1, -1, -1, -1, -1, -1, -1
## $ tilt_1                      <dbl> -1, -1, -1, -1, -1, -1, -1
## $ tilt_2                      <dbl> -1, -1, -1, -1, -1, -1, -1
## $ tilt_3                      <dbl> -1, -1, -1, -1, -1, -1, -1
## $ module_manufacturer_1       <chr> "-1", "Hanwha Q CELLS", "-
## $ module_model_1              <chr> "-1", "Q.PEAK DUO-G7 325",
## $ module_quantity_1           <dbl> -1, -1, -1, -1, -1, -1, -1
## $ module_manufacturer_2       <chr> "-1", "-1", "-1", "-1", "-
## $ module_model_2              <chr> "-1", "-1", "-1", "-1", "-
## $ module_quantity_2           <dbl> -1, -1, -1, -1, -1, -1, -1
## $ module_manufacturer_3       <chr> "-1", "-1", "-1", "-1", "-
## $ module_model_3              <chr> "-1", "-1", "-1", "-1", "-
## $ module_quantity_3           <dbl> -1, -1, -1, -1, -1, -1, -1
## $ additional_modules          <dbl> -1, -1, -1, -1, -1, -1, -1
## $ technology_module_1         <chr> "-1", "Monocrystalline", "
## $ technology_module_2         <chr> "-1", "-1", "-1", "-1", "-
## $ technology_module_3         <chr> "-1", "-1", "-1", "-1", "-
## $ BIPV_module_1               <dbl> -1, 0, -1, 0, 0, -1, -1, 0
## $ BIPV_module_2               <dbl> -1, -1, -1, -1, -1, -1, -1
## $ BIPV_module_3               <dbl> -1, -1, -1, -1, -1, -1, -1
## $ bifacial_module_1           <dbl> -1, 0, -1, 0, 0, -1, -1, 0
## $ bifacial_module_2           <dbl> -1, -1, -1, -1, -1, -1, -1
## $ bifacial_module_3           <dbl> -1, -1, -1, -1, -1, -1, -1
## $ nameplate_capacity_module_1 <dbl> -1, 325, -1, 320, 230, -1,
## $ nameplate_capacity_module_2 <dbl> -1, -1, -1, -1, -1, -1, -1
## $ nameplate_capacity_module_3 <dbl> -1, -1, -1, -1, -1, -1, -1
## $ efficiency_module_1         <dbl> -1.00000000, 0.19938650, -
## $ efficiency_module_2         <dbl> -1, -1, -1, -1, -1, -1, -1
## $ efficiency_module_3         <dbl> -1, -1, -1, -1, -1, -1, -1
## $ inverter_manufacturer_1     <chr> "-1", "Enphase Energy Inc.
## $ inverter_model_1            <chr> "-1", "IQ7-60-2-US [240V]"
## $ inverter_quantity_1         <dbl> -1, -1, -1, -1, -1, -1, -1
## $ inverter_manufacturer_2     <chr> "-1", "-1", "-1", "-1", "-
## $ inverter_model_2            <chr> "-1", "-1", "-1", "-1", "-
## $ inverter_quantity_2         <dbl> -1, -1, -1, -1, -1, -1, -1
## $ inverter_manufacturer_3     <chr> "-1", "-1", "-1", "-1", "-
```

```
## $ inverter_model_3           <chr> "-1", "-1", "-1", "-1", "-
## $ inverter_quantity_3        <dbl> -1, -1, -1, -1, -1, -1, -1
## $ additional_inverters       <dbl> -1, -1, -1, -1, -1, -1, -1
## $ micro_inverter_1           <dbl> -1, 1, -1, 0, -1, -1, -1,
## $ micro_inverter_2           <dbl> -1, -1, -1, -1, -1, -1, -1
## $ micro_inverter_3           <dbl> -1, -1, -1, -1, -1, -1, -1
## $ built_in_meter_inverter_1  <dbl> -1, 0, -1, 0, -1, -1, -1,
## $ built_in_meter_inverter_2  <dbl> -1, -1, -1, -1, -1, -1, -1
## $ built_in_meter_inverter_3  <dbl> -1, -1, -1, -1, -1, -1, -1
## $ output_capacity_inverter_1 <dbl> -1.00, 0.24, -1.00, -1.00,
## $ output_capacity_inverter_2 <dbl> -1, -1, -1, -1, -1, -1, -1
## $ output_capacity_inverter_3 <dbl> -1, -1, -1, -1, -1, -1, -1
## $ DC_optimizer               <dbl> -1, 0, -1, 1, -1, -1, -1,
## $ inverter_loading_ratio     <dbl> -1.0000000, 1.3541667, -1.
## $ battery_manufacturer       <chr> "-1", "-1", "-1", "-1", "-
## $ battery_model              <chr> "-1", "-1", "-1", "-1", "-
## $ battery_rated_capacity_kW  <dbl> -1, -1, -1, -1, -1, -1, -1
## $ battery_rated_capacity_kWh <dbl> -1, -1, -1, -1, -1, -1, -1
## $ battery_price              <dbl> -1, -1, -1, -1, -1, -1, -1
## $ technology_type            <chr> "pv-only", "pv-only", "pv-
## $ extensions_multiphase_id   <chr> "-1", "-1", "-1", "-1", "-
```

That's a lot of -1s. If we look at the documentation, we see that missing data is coded as -1!

When is an observation missing just too much data? It's probably fine if we don't have any azimuth data. On the other hand, if we don't have a date of purchase, it'll be literally impossible for us to put the observation into a graph. We probably also want to subset to observations where the manufacturer is available, the quantity purchased is available, and which include only solar panels:

```
# subset - missing years are encoded as NaT
tts_sub = tts_raw |> filter(!near(efficiency_module_1, -1, 0.01))
                     filter(technology_type == "pv-only",
                            module_manufacturer_1 != "-1",
                            module_quantity_1 > -1,
                            installation_date != "NaT") |>
  mutate(year = year(dmy(installation_date)))

glimpse(tts_sub)
```

```
## Rows: 1,795,710
## Columns: 82
## $ data_provider_1             <chr> "CPS Energy", "CPS Energy"
## $ data_provider_2             <chr> "-1", "-1", "-1", "-1", "-
## $ system_ID_1                 <chr> "4874", "18995", "1243", "
## $ system_ID_2                 <chr> "-1", "-1", "-1", "-1", "-
## $ installation_date           <chr> "16-Mar-2016", "11-Apr-201
## $ PV_system_size_DC           <dbl> 8.82, 9.24, 3.45, 7.00, 10
## $ total_installed_price       <dbl> 25774.60, 34107.00, 20503.
## $ rebate_or_grant             <dbl> 12428.21, 2500.00, 6076.80
## $ customer_segment            <chr> "RES", "RES", "RES", "RES"
## $ expansion_system            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ multiple_phase_system       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ TTS_link_ID                 <chr> "-1", "-1", "-1", "-1", "-
## $ new_construction            <dbl> -1, -1, -1, -1, -1, -1, -1
## $ tracking                    <dbl> -1, -1, -1, -1, -1, -1, -1
## $ ground_mounted              <dbl> -1, -1, -1, -1, -1, -1, -1
## $ zip_code                    <chr> "78261", "78263", "78023",
## $ city                        <chr> "San Antonio", "San Antoni
## $ state                       <chr> "TX", "TX", "TX", "TX", "T
## $ utility_service_territory   <chr> "CPS Energy", "CPS Energy"
## $ third_party_owned           <dbl> -1, -1, -1, -1, -1, -1, -1
## $ installer_name              <chr> "Advanced Solar Electric",
## $ self_installed              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ azimuth_1                   <dbl> -1, -1, -1, -1, -1, -1, -1
## $ azimuth_2                   <dbl> -1, -1, -1, -1, -1, -1, -1
## $ azimuth_3                   <dbl> -1, -1, -1, -1, -1, -1, -1
## $ tilt_1                      <dbl> -1, -1, -1, -1, -1, -1, -1
## $ tilt_2                      <dbl> -1, -1, -1, -1, -1, -1, -1
## $ tilt_3                      <dbl> -1, -1, -1, -1, -1, -1, -1
```

```
## $ module_manufacturer_1       <chr> "ET Solar Industry", "SANY
## $ module_model_1              <chr> "ET-P660245B", "VBHN330SA1
## $ module_quantity_1           <dbl> 36, 28, 15, 28, 37, 35, 24
## $ module_manufacturer_2       <chr> "-1", "-1", "-1", "-1", "-
## $ module_model_2              <chr> "-1", "-1", "-1", "-1", "-
## $ module_quantity_2           <dbl> -1, -1, -1, -1, -1, -1, -1
## $ module_manufacturer_3       <chr> "-1", "-1", "-1", "-1", "-
## $ module_model_3              <chr> "-1", "-1", "-1", "-1", "-
## $ module_quantity_3           <dbl> -1, -1, -1, -1, -1, -1, -1
## $ additional_modules          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ technology_module_1         <chr> "Polycrystalline", "Monocr
## $ technology_module_2         <chr> "-1", "-1", "-1", "-1", "-
## $ technology_module_3         <chr> "-1", "-1", "-1", "-1", "-
## $ BIPV_module_1               <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ BIPV_module_2               <dbl> -1, -1, -1, -1, -1, -1, -1
## $ BIPV_module_3               <dbl> -1, -1, -1, -1, -1, -1, -1
## $ bifacial_module_1           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ bifacial_module_2           <dbl> -1, -1, -1, -1, -1, -1, -1
## $ bifacial_module_3           <dbl> -1, -1, -1, -1, -1, -1, -1
## $ nameplate_capacity_module_1 <dbl> 245, 330, 230, 250, 280, 3
## $ nameplate_capacity_module_2 <dbl> -1, -1, -1, -1, -1, -1, -1
## $ nameplate_capacity_module_3 <dbl> -1, -1, -1, -1, -1, -1, -1
## $ efficiency_module_1         <dbl> 0.1520677, 0.2062500, 0.14
## $ efficiency_module_2         <dbl> -1, -1, -1, -1, -1, -1, -1
## $ efficiency_module_3         <dbl> -1, -1, -1, -1, -1, -1, -1
## $ inverter_manufacturer_1     <chr> "SolarEdge Technologies Lt
## $ inverter_model_1            <chr> "SE7600A-US [240V]", "-1",
## $ inverter_quantity_1         <dbl> 1, 1, 15, 1, 37, 1, 1, 1,
## $ inverter_manufacturer_2     <chr> "-1", "-1", "-1", "-1", "-
## $ inverter_model_2            <chr> "-1", "-1", "-1", "-1", "-
## $ inverter_quantity_2         <dbl> -1, -1, -1, -1, -1, -1, -1
## $ inverter_manufacturer_3     <chr> "-1", "-1", "-1", "-1", "-
## $ inverter_model_3            <chr> "-1", "-1", "-1", "-1", "-
## $ inverter_quantity_3         <dbl> -1, -1, -1, -1, -1, -1, -1
## $ additional_inverters        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ micro_inverter_1            <dbl> 0, -1, -1, 0, 1, 0, 0, 0,
## $ micro_inverter_2            <dbl> -1, -1, -1, -1, -1, -1, -1
## $ micro_inverter_3            <dbl> -1, -1, -1, -1, -1, -1, -1
## $ built_in_meter_inverter_1   <dbl> 1, -1, -1, 1, 0, 0, 1, 0,
## $ built_in_meter_inverter_2   <dbl> -1, -1, -1, -1, -1, -1, -1
```

```
## $ built_in_meter_inverter_3   <dbl> -1, -1, -1, -1, -1, -1, -1
## $ output_capacity_inverter_1  <dbl> 7.625, -1.000, -1.000, 6.0
## $ output_capacity_inverter_2  <dbl> -1, -1, -1, -1, -1, -1, -1
## $ output_capacity_inverter_3  <dbl> -1, -1, -1, -1, -1, -1, -1
## $ DC_optimizer                <dbl> 1, 0, 0, 0, 0, 1, 0, 1, 1,
## $ inverter_loading_ratio      <dbl> 1.156721, -1.000000, -1.00
## $ battery_manufacturer        <chr> "-1", "-1", "-1", "-1", "-
## $ battery_model               <chr> "-1", "-1", "-1", "-1", "-
## $ battery_rated_capacity_kW   <dbl> -1, -1, -1, -1, -1, -1, -1
## $ battery_rated_capacity_kWh  <dbl> -1, -1, -1, -1, -1, -1, -1
## $ battery_price               <dbl> -1, -1, -1, -1, -1, -1, -1
## $ technology_type             <chr> "pv-only", "pv-only", "pv-
## $ extensions_multiphase_id    <chr> "-1", "-1", "-1", "-1", "-
## $ year                        <int> 2016, 2019, 2012, 2014, 20
```

Looks a lot better. Some of these manufacturers are a little weird though: Sanyo Electric and CSI Solar are big-name manufacturers, ET Solar is really not.

We could look up a list of big solar manufacturers and subset to that list, but that method would be a little ad hoc. Instead, we can go and calculate yearly market shares. Then we can use a market share threshold to select big manufacturers over time.

```
# manuf market shares
shares = tts_sub |> group_by(year) |>
  mutate(market_sales = sum(module_quantity_1, na.rm=TRUE)) |>
  group_by(module_manufacturer_1, year, market_sales) |>
  summarize(manuf_sales = sum(module_quantity_1, na.rm = TRUE),
            shares = manuf_sales / first(market_sales),
            .groups = "drop") |>
```

```
  select(year, module_manufacturer_1, manuf_sales, market_sales,
head(shares)
```

```
## # A tibble: 6 × 5
##    year module_manufacturer_1 manuf_sales market_sales    shar
##   <int> <chr>                       <dbl>        <dbl>     <db
## 1  2011 1Soltech                      697      1662436 0.00041
## 2  2012 1Soltech                     1036      2074577 0.00049
## 3  2013 1Soltech                      581      2816324 0.00020
## 4  2014 1Soltech                      157      3551933 0.00004
## 5  2004 APOS Energy                     8        62116 0.00012
## 6  2005 APOS Energy                    14        64579 0.00021
```

```
# subset to firms with 10% market share, get list of names
top_firm_data = shares |> filter(shares >= 0.10, year >= 2010) |>
  group_by(year) |> mutate(top_firm_share = sum(shares)) |> ungro
top_firms = top_firm_data |> select(module_manufacturer_1) |> dis
flist = as.list(top_firms$module_manufacturer_1)
print(flist)
```

```
## [[1]]
## [1] "CSI Solar Co., Ltd."
##
## [[2]]
## [1] "First Solar, Inc."
##
## [[3]]
## [1] "Hanwha Q CELLS"
##
## [[4]]
## [1] "LG Electronics Inc."
##
## [[5]]
## [1] "LONGi Green Energy Technology Co., Ltd."
##
## [[6]]
## [1] "REC Solar"
##
## [[7]]
## [1] "Sharp"
```

```
##
## [[8]]
## [1] "SolarWorld"
##
## [[9]]
## [1] "SunPower"
##
## [[10]]
## [1] "Suntech Power"
##
## [[11]]
## [1] "Trina Solar"
##
## [[12]]
## [1] "Yingli Energy (China)"
```

CSI made our cut, but Sanyo didn't! Good thing we didn't trust the Panasonic name - they must not be a big player in the US.

Okay. Next: subset to firms only in the top firm list and find their top efficiency module per year:

```
# subset tts_sub to firm list, find top efficiency module per yea
module_efficiency = tts_sub |> filter(module_manufacturer_1 %in%
  group_by(module_manufacturer_1, technology_module_1, year) |>
  mutate(manuf_max_eff = max(efficiency_module_1)) |>
  group_by(year, module_manufacturer_1, module_model_1,
           manuf_max_eff, technology_module_1) |>
  summarize(price = mean(total_installed_price),
            efficiency = mean(efficiency_module_1),
            .groups = "drop")
# generate year-manufacturer-technology-max efficiency df
manuf_tech_eff = module_efficiency |>
  select(year, module_manufacturer_1, technology_module_1, manuf_
  distinct() |>
```

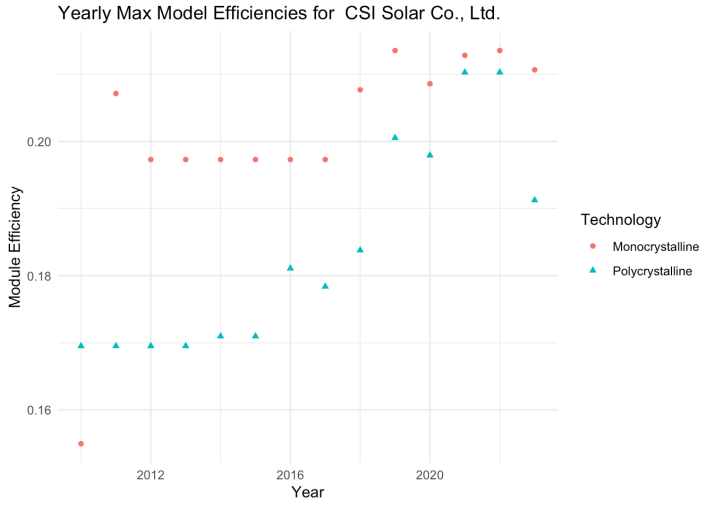
```
  filter(year >= 2010, str_detect(technology_module_1,regex("Monc
head(manuf_tech_eff)
```

```
## # A tibble: 6 × 4
##    year module_manufacturer_1 technology_module_1 manuf_max_ef
##   <int> <chr>                <chr>                       <dbl
## 1  2010 CSI Solar Co., Ltd.  Polycrystalline             0.17
## 2  2010 CSI Solar Co., Ltd.  Monocrystalline             0.15
## 3  2010 Hanwha Q CELLS       Polycrystalline             0.16
## 4  2010 REC Solar            Polycrystalline             0.14
## 5  2010 REC Solar            Monocrystalline             0.19
## 6  2010 Sharp                Polycrystalline             0.14
```

We're almost there. Now we know the max panel efficiency for each manufacturer-year-technology group in our data. All that's left is to plot that data!

```
# create ladder plot for 2010
years = 2010:2022
manuf_tech_eff$technology_module_1 = as.factor(manuf_tech_eff$tec

plot_maxes <- function(z) ggplot(manuf_tech_eff[manuf_tech_eff$mc
                                 aes(x=year,y=manuf_max_eff, col=
                                     shape=technology_module_1))
                               labs(title=paste("Yearly Max Model
                                    x="Year",y="Module Efficiency"
                                    col="Technology") +
                               geom_point() +
                               theme_minimal()
p <- lapply(flist,plot_maxes)
p[[1]]
```



Presentations

Your final project will consist of the code you wrote and two presentations you created. The first presentation will be a 5-slide, 5-minute presentation you deliver during our scheduled exam time. The second presentation will be a final work product. While you'll never present it per se, the grader and I will use it to assess your final project.

I'm going to be fairly prescriptive about what's in the in-class presentation. I expect 5 slides in total: an introduction slide, which contains your research question and a brief motivation as to why we should care about your question; a data slide, which describes the datasets you're using; 2 slides for findings - your discretion as to what's included here, but this would be a nice place to show a graph and/or a regression table; and one slide to

wrap up. Think about this as being similar to a presentation you'd give in a workplace. Remember not to use too many words on your slides!

The 20-slide presentation, your final work product, can be more involved. In it, I'd like you to be more explicit about your project: show us what the data looks like, what kind of cleaning you did, what the assumptions and drawbacks are to your regression, and so on. Think of it as an essay in slide deck format.

I'd like both presentations to be done in RMarkdown. You can use my slides from this semester as a template.