PA 446

Coding for Civic Data Applications

Will be starting at 6:05pm

Class #8

Logistics

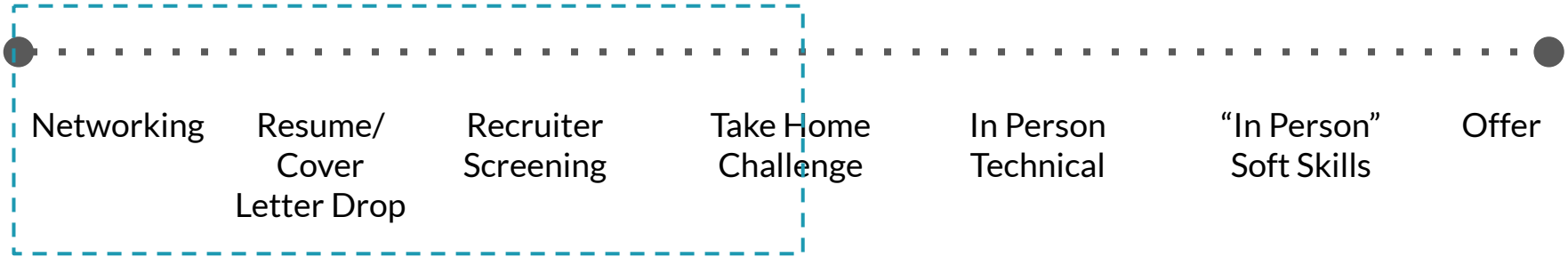
Course Logistics

- HW 4
  - Due 10/20

Class #8

Content: Story-telling

Focus Last Week



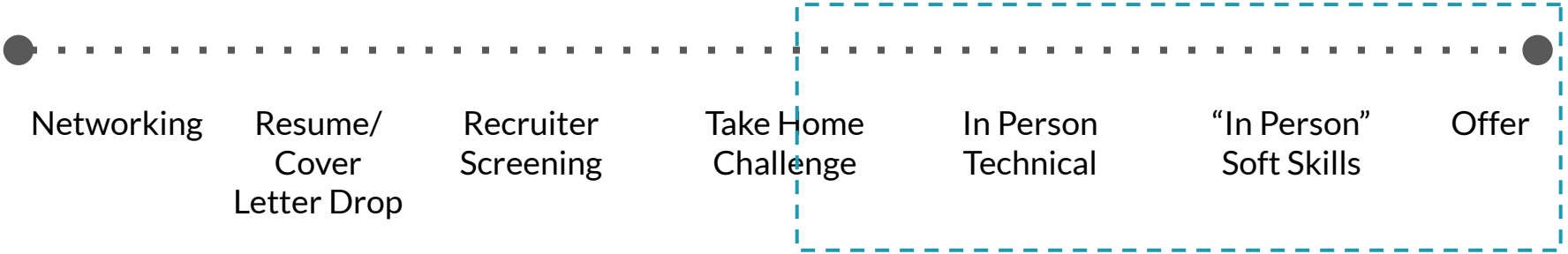
1. Review of the first 4 steps of data science job application process
2. Introduction to SQL

Pushing Back on Salary Question

Example Response

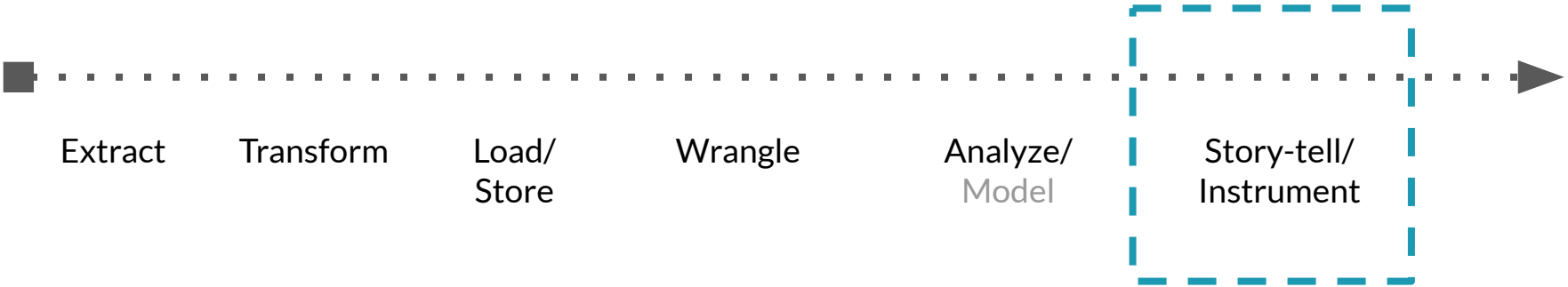
I definitely understand that this is important to discuss and I will be happy to have this conversation as a later time. My salary expectations will depend on what my roles and responsibilities will ultimately look like. To that end, I would like to go further in this interview process and learn more about it from the hiring manager and other data scientist.

Focus This Week



1. Wrap up the take-home challenge with story-telling and presentation with data
2. Review of the final 3 steps

Additional Focus This Week



Where We Been

Take Home **Analysis** Framework

- Data cleaning
- Minimal feature selection
- Impute missing values
- Create a modeling pipeline / analysis code
  - If modeling, training with a couple of classifiers
  - If modeling, tune hyperparameters
- Visualize and package findings

Where We Been

Take Home **Analysis** Framework

- Data cleaning
- Minimal feature selection
- Impute missing values
- Create a modeling pipeline / analysis code
  - If modeling, training with a couple of classifiers
  - If modeling, tune hyperparameters
- Visualize and package findings

After your first pass at analysis, it is time to think about storytelling

Take Home Challenge

Storytelling Framework

Who: understand your audience

Goal: what are you trying to convince your audience

How: what are you going to show your audience to convince them

Who

Understand Your Audience

Who

Key Questions

1. Who is your audience's boss
2. Bias for action vs consensus driven
3. What is your audience's level of technicality

Who

Discuss Mayor LightFoot

1. Who is her boss
2. Bias for action vs consensus driven
3. What is her level of technicality

Who

Key Questions: So What

1. Who is your audience's boss: **focus of your analysis on what the boss cares about**
2. Bias for action vs consensus driven: **should you have a bias for action or triple check your analysis**
3. What is your audience's level of technicality: **how technical can you be in your presentation**

Storytelling Framework

What Does This Imply?

Understand Mayor Lightfoot

1. Boss: Chicago voters
    a. Your looked at gender + race pay disparity across the city's 5 largest departments. What might be of interest to voters?
2. Bias for action vs consensus: consensus
    a. You have to triple check your work - what should you focus on?
3. Level of technicality: not very
    a. Don't brag about your linear regression's R-square

Goal

What are you trying to convince your audience

Goal

Key Questions

1. Findings that mattered
2. Remaining uncertainties
3. Additional work or assumptions needed

Goal

Discuss: Chicago Salaries Data

1. Findings that mattered:
2. Remaining uncertainties:
3. Additional work or assumptions needed:

Goal

Key Questions: So What

1. What are the important findings: **importance in consulting = what is most actionable. Can vary in other industries**
2. What are remaining uncertainties: **among the important findings, which are you not sure about**
3. Additional work or assumptions needed: **figure out if/how to address these uncertainties**

Goal

Discuss: Chicago Salaries Data

1. Important findings
    a. **Men made more than women**
    b. **Salary ranked by race: API, White, Hispanic, African American (story gets nuanced at the dept level)**
2. Remaining uncertainties:
    a. **Imputed gender and race**
    b. **Not controlling for job title**
    c. **Low R-square**

Goal

Discuss: Chicago Salaries Data

3. Additional work or assumptions needed:
    a. Confirm gender/race distribution with available data
    b. **Really low R-square - figure out how to control for job title**

Goal

Low R-Square: Discuss

Controlling for job title. Suggestions?

-

Goal

Low R-Square

[to the code]

Brief Tangent on Variance and Bias

Not required on HW4

Terminology

Bias: how accurate is your model

Variance: how jumpy is your model's predictions and how generalizable is your model to the population at large

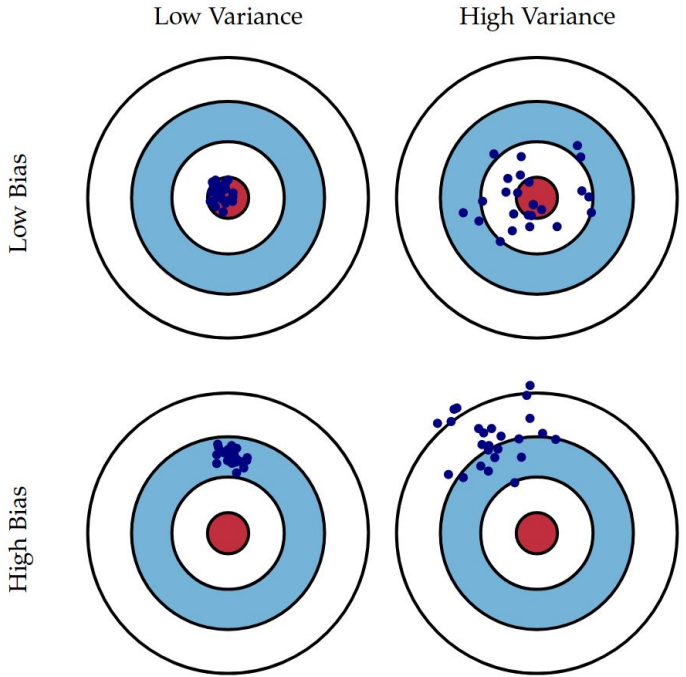


Fig. 1 Graphical illustration of bias and variance.

Brief Tangent on Variance and Bias

Bias and Variance, in our case

For linear models

a. Bias: approximated by R-square
b. Variance: there's a formula, but more important to remember that if you have >20 IV's, the grader will suspect that your model might have high variance

Goal

Key Questions

1. Findings that mattered: **white and male city workers tend to make more**
2. Remaining uncertainties: **had to impute race, and initial model had low R-square**
3. Additional work or assumptions needed: **confirm accuracy of imputation and improve linear model**

Goal

What are you trying to convince your audience

term	estimate	p.value
(Intercept)	81839.2	0.0000
final_race_twoblack	2673.5	0.0038
final_race_twohispanic	494.1	0.5194
final_race_twowhite	2999.1	0.0001
`Job Titles`POLICE OFFICER (ASSIGNED AS DETECTIVE)	13855.2	0.0000
`Job Titles`SERGEANT	36610.0	0.0000
genderM	1135.5	0.0000

Story

Within police's most common jobs, men makes more than women. Hispanic and API officers tend to make less than White and African American officers

Goal

What are you trying to convince your audience

term	estimate	p.value
(Intercept)	105381.448	0.000
final_race_twoblack	-2802.655	0.393
final_race_twohispanic	-4946.485	0.116
final_race_twowhite	-4734.577	0.131
`Job Titles`FIREFIGHTER-EMT (RECRUIT)	-30538.206	0.000
genderM	1770.758	0.060

Story

Within the fire department's most common jobs, there are no significant pay disparities along the lines of race and gender

How

What are you going to show your audience to convince them?

How

Framework

- Data Reinforcing Stories > Stories Reinforcing Data
- For Non-technical Stakeholders, Appeal on a Personal Level
- Reduce the number of ‘moving pieces’

Data Reinforcing Stories > Stories Reinforcing Data

You are a data scientist, but not at the expense of storytelling

- Too much data without enough storytelling lacks proper context, personal connection, and narrative coherence: *"why should I care?"* **More common amongst data scientists**
- Too much storytelling without enough data lacks credibility, extensibility, and concrete policy proposals: *"why should I believe you?"*

Data Reinforcing Stories > Stories Reinforcing Data

Too Much Data

term	estimate	p.value
(Intercept)	81839.2	0.0000
final_race_twoblack	2673.5	0.0038
final_race_twohispanic	494.1	0.5194
final_race_twowhite	2999.1	0.0001
`Job Titles`POLICE OFFICER (ASSIGNED AS DETECTIVE)	13855.2	0.0000
`Job Titles`SERGEANT	36610.0	0.0000
genderM	1135.5	0.0000

term	estimate	p.value
(Intercept)	105381.448	0.000
final_race_twoblack	-2802.655	0.393
final_race_twohispanic	-4946.485	0.116
final_race_twowhite	-4734.577	0.131
`Job Titles`FIREFIGHTER-EMT (RECRUIT)	-30538.206	0.000
genderM	1770.758	0.060

Data Reinforcing Stories > Stories Reinforcing Data

Better, but the Story is still Enforcing the Data

term	estimate	p.value
(Intercept)	81839.2	0.0000
final_race_twoblack	2673.5	0.0038
final_race_twohispanic	494.1	0.5194
final_race_twowhite	2999.1	0.0001
`Job Titles`POLICE OFFICER (ASSIGNED AS DETECTIVE)	13855.2	0.0000
`Job Titles`SERGEANT	36610.0	0.0000
genderM	1135.5	0.0000

Story

Within police's most common jobs, men makes more than women. Hispanic and API officers tend to make less than White and African American officers

Data Reinforcing Stories > Stories Reinforcing Data

How Can You Make the Data Enforce the Story?

term	estimate	p.value
(Intercept)	81839.2	0.0000
final_race_twoblack	2673.5	0.0038
final_race_twohispanic	494.1	0.5194
final_race_twowhite	2999.1	0.0001
`Job Titles`POLICE OFFICER (ASSIGNED AS DETECTIVE)	13855.2	0.0000
`Job Titles`SERGEANT	36610.0	0.0000
genderM	1135.5	0.0000

Story

Within police's most common jobs, men makes more than women. Hispanic and API officers tend to make less than White and African American officers

Data Reinforcing Stories > Stories Reinforcing Data

How Can You Make the Data Enforce the Story?

term	estimate	p.value
(Intercept)	81839.2	0.0000
final_race_twoblack	2673.5	0.0038
final_race_twohispanic	494.1	0.5194
final_race_twowhite	2999.1	0.0001
`Job Titles`POLICE OFFICER (ASSIGNED AS DETECTIVE)	13855.2	0.0000
`Job Titles`SERGEANT	36610.0	0.0000
genderM	1135.5	0.0000

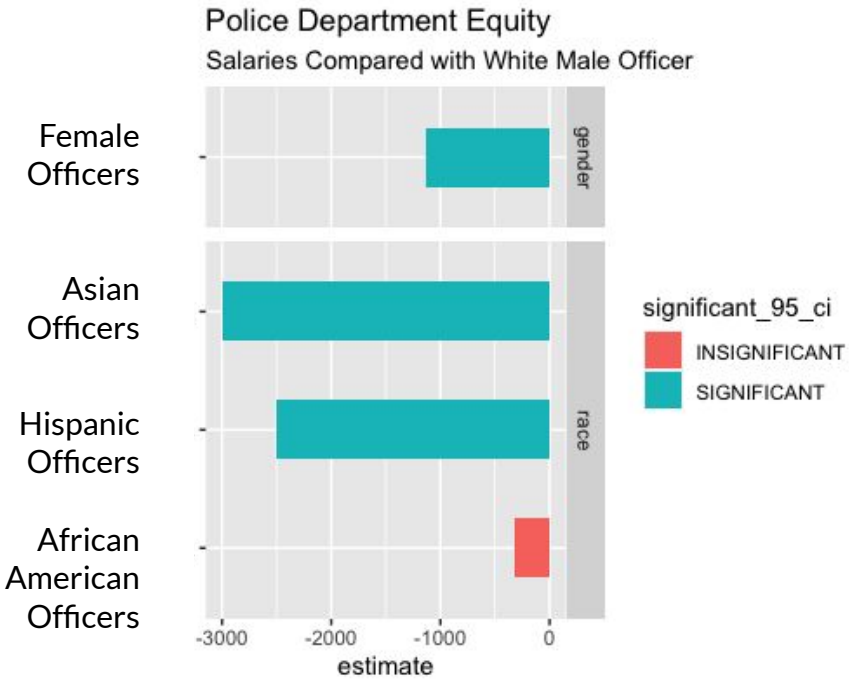
[to coding]

15-minute Break

Be Back at 7:45pm

Data Reinforcing Stories > Stories Reinforcing Data

How Can You Make the Data Enforce the Story?



Difference in Average Annual Salary (USD)

Story

Within police's most common jobs, men makes more than women

Hispanic and API officers tend to make less than White and African American officers

Appeal on a Personal Level

Especially for Non-technical Stakeholders

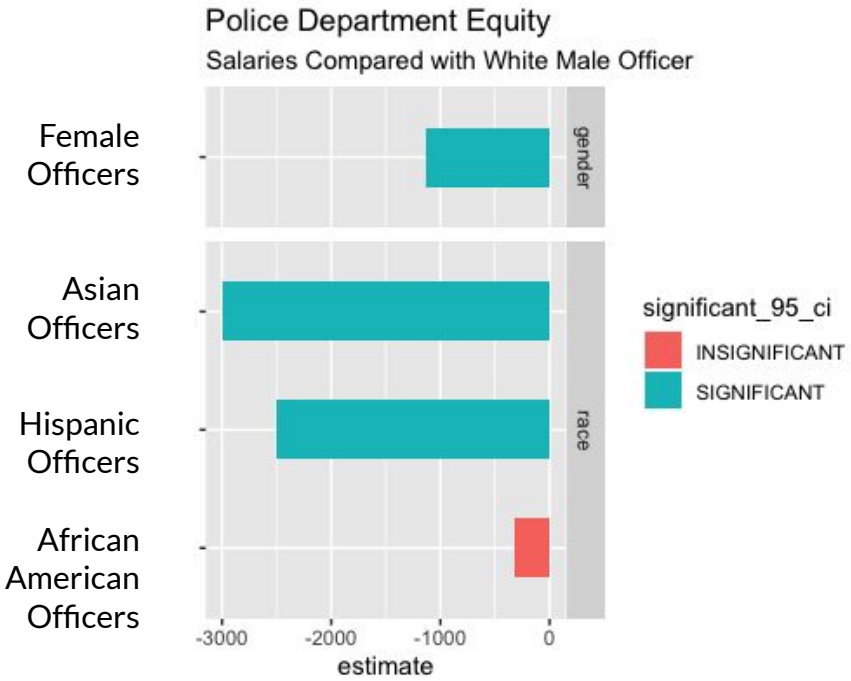
Personal isn't necessarily emotional. Just means grounded in examples, ideally examples that involves people.

Start at 0:20
https://www.npr.org/2021/03/12/976465414/the-even-more-minimum-wage

https://www.cnbc.com/2021/01/26/democrats-reintroduce-15-minimum-wage-bill-with-unified-control-of-congress.html

Appeal on a Personal Level

Especially for Non-technical Stakeholders

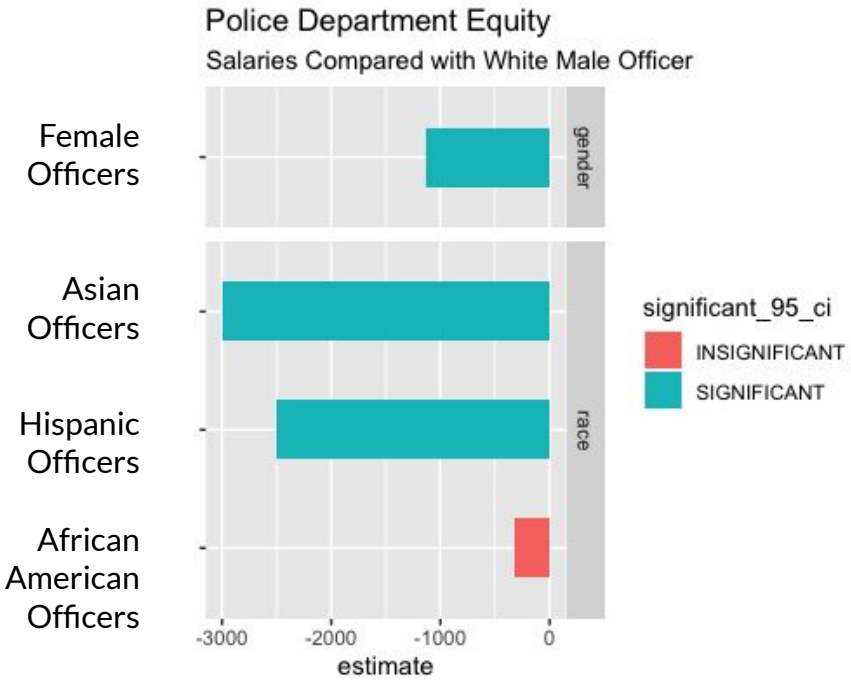


Difference in Average Annual Salary (USD)

How can we make this more personal?

Appeal on a Personal Level

Especially for Non-technical Stakeholders



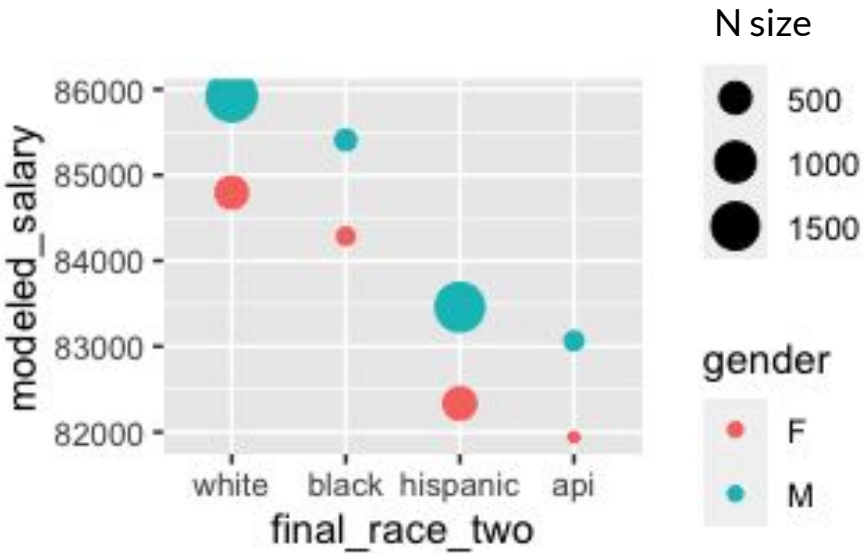
Difference in Average Annual Salary (USD)

Salary numbers are relative, no absolute

Missing interaction between gender and race

Appeal on a Personal Level

Especially for Non-technical Stakeholders



Story

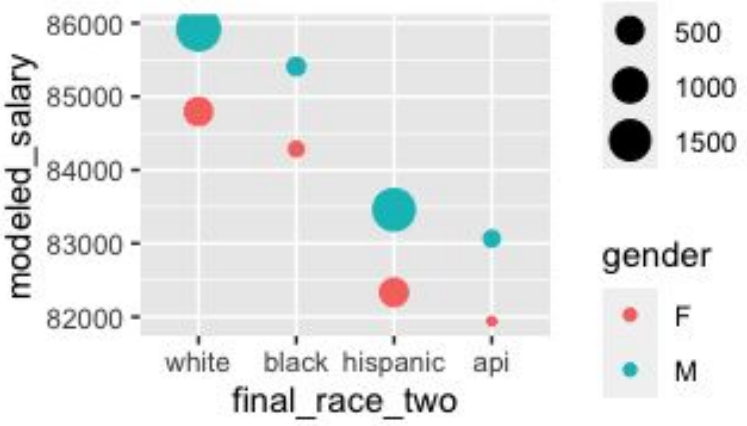
Within police's most common jobs, men makes more than women

Hispanic and API officers tend to make less than White and African American officers

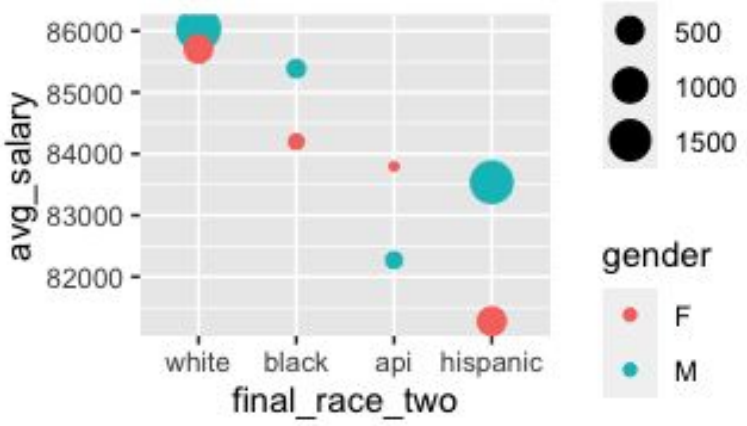
Appeal on a Personal Level

Modeled Results vs Actuals

Modeled Salaries



Actual



Reduce the number of moving pieces

- The number of **takeaways** in your presentation
- The types of **data cuts** in your presentation: types of graphs, ways of filtering your results
- The **formatting** of your graphics: colors and fonts

Reduce the number of moving pieces

- The number of **takeaways** in your presentation
- The types of **data cuts** in your presentation: types of graphs, ways of filtering your results
- The **formatting** of your graphics: colors and fonts

Leading NBA scorers by zone: interesting and good

Reduce the number of moving pieces

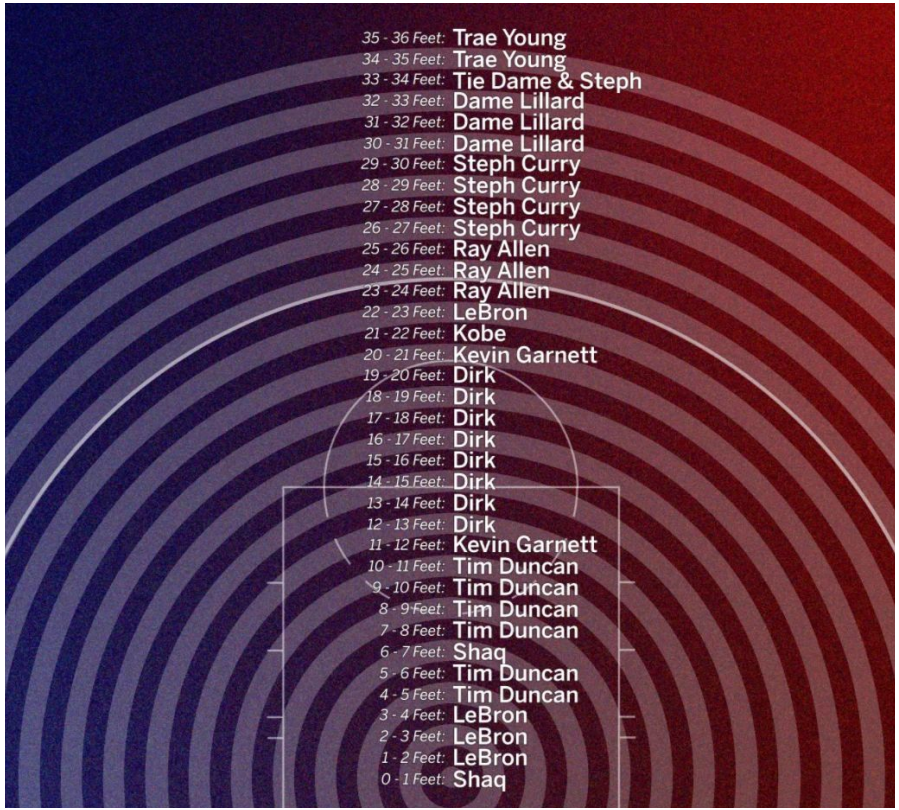
- The number of **takeaways** in your presentation
- The types of **data cuts** in your presentation: types of graphs, ways of filtering your results
- The **formatting** of your graphics: colors and fonts

Pop quiz: what takeaways do you have from that image?

Reduce the number of moving pieces

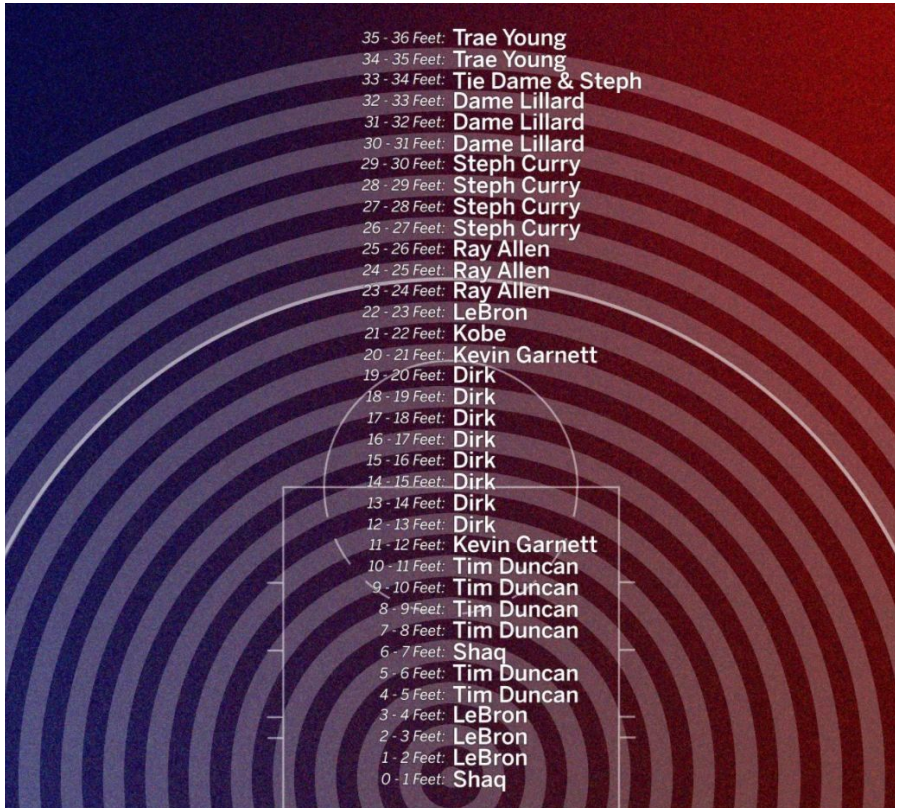
- The number of **takeaways** in your presentation
- The types of **data cuts** in your presentation: types of graphs, ways of filtering your results
- The **formatting** of your graphics: colors and fonts

Leading NBA scorers by zone: great



Reduce the number of moving pieces

- The number of **takeaways** in your presentation: **clearer picture of who is good where**
- The types of **data cuts** in your presentation: types of graphs, ways of filtering your results: **just distance from the basket**
- The **formatting** of your graphics: colors and fonts: **just names, no head shots**



Take Home

Wrap Up

Review Your Work!

Part of the reason to not work till the last minute

- First pass by yourself
- If possible, have peers, especially other data scientists review your work after you submit for feedback

Summary: Linear Processes Are Rare

Data Science Is Iterative