**Phoenix** hcbb

Assignment: Option 2

1. **Physician & Other Supplier Payments *-- 2015, 2016 and 2017 (pickled)******“Provider Utilization File” - slides pgs. 10-20*** <https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier2017>Downloaded in class\_notebook pickles\_and\_chunks  
   Only saved the resulting pickle file to data in HCBB repo
2. **Detailed Data Hospital Outpatient**<https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Outpatient>
3. **Detailed Data APC to CPT/HCPCS crosswalk, Addendum B – January 2020 (correction files aren't necessary)**<https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/HospitalOutpatientPPS/Addendum-A-and-Addendum-B-Updates>
4. **Zip Code to CBSA**  
   <https://www.huduser.gov/portal/datasets/usps_crosswalk.html>
5. **Other** [https://Data.CMS.gov](https://data.cms.gov/)

**Goals:**

**Measure how payments and counts change over time**

**“Counts” = count of Beneficiaries and count of Services**

***DATA: #1, for 2015, 2016 and 2017***

***[Average Medicare Allowed Amount= How much did the Provider receives (from Medicare + any patient responsibility)]***

Download 2015-2017 data (or earlier)

Join columns for Physician/Other Supplier *(join 2015, 2016, 2017 - on selected columns)*

NPI / HCPCS / Place of Service *(Use these as primary key)*

Which procedures had the largest change in Average payment?

Utilization?

**Option 2 Stretch Goals:**

**Include Hospitals in analysis**  *Possibly only 2016-2017 (2015 has diff. codes)*

* Join columns for Hospital
* Provider ID / APC

Remember APCs changed in 2016, so you’ll need to convert old APCs from 2015 and before to the new APCs

#### **Columns Using:**[**¶**](http://localhost:8888/notebooks/Desktop/nss_data_analytics/projects/healthcare-bluebook-phoenix/notebooks/hcbb.ipynb#Let's-keep-the-following-columns:)

* **National Provider Identifier** -- int64
* **Last Name/Organization Name of the Provider** -- l object
* **Entity Type of the Provider** -- l object
* **City of the Provider** -- object
* **Zip Code of the Provider** -- int64
* **State Code of the Provider** -- object
* **Provider Type** -- object
* **Place of Service** -- object
* **HCPCS Code** -- object
* **HCPCS Description** -- object
* **Number of Services** -- float64
* **Number of Medicare Beneficiaries** -- int64
* **Number of Distinct Medicare Beneficiary/Per Day Services** -- int64
* **Average Medicare Allowed Amount** -- float64
* **ADD: Year** (in each df on import)

**Columns Dropped (with 30M rows, looking to drop as much as possible for efficiency)**

* First Name of the Provider -- 976 non-null object
  + NPI used to identify
* Middle initial of provider
  + NPI used to identify
* Gender of the provider
  + No analysis expected by gender
* Street Address 1 of the Provider -- 1000 non-null object
  + Keeping higher level location (state, city, zip)
* Street address 2 of the provider
  + Keeping higher level location (state, city, zip)
* Average Submitted Charge Amount -- 1000 non-null float64
  + HCBB communicated this is not used
* Average Medicare Payment Amount -- 1000 non-null float64
  + HCBB communicated this is only medicare contribution, will use total for analysis
* Average Medicare Standardized Amount -- 1000 non-null float64
  + HCBB communicated “Allowed Amount” should be used
* Medicare Participation Indicator -- 1000 non-null object \*\*
  + Via visualization tool, very few observations were “N”, 99% were “Y”
  + Based on meta-data, this indicator is just participation in medicare, should not affect analysis
* *Credentials of the Provider -- 936 non-null object MAYBE USE LATER?*
* Country Code of the Provider -- 1000 non-null object
  + 99% are USA.
* HCPCS Drug Indicator -- 1000 non-null object
  + Only indicates that drug pricing included in a separate document. Not needed for this analysis

General formatting:

* Lowercase
* Underscores

*\*\* medicare\_participation\_indicator –*

*There are only 3,208 “N” in 2017. Not significant amount.*

*Identifies whether the provider participates in Medicare and/or accepts assignment of Medicare allowed amounts. The value will be ‘Y’ for any provider that had at least one claim identifying the provider as participating in Medicare or accepting assignment of Medicare allowed amounts within HCPCS code and place of service. A non-participating provider may elect to accept Medicare allowed amounts for some services and not accept Medicare allowed amounts for other services.*

Approach:

Break each step into teams.

**Due by end of day Friday 5/29 - DONE**

Initial EDA: Validate data answers question (completed)

ETL: extract, organize, wall time tests, handled nulls, pickled, combined datasets - did as a group, to ensure consistency of data we’re each using. (completed)

Divided into two sub-teams to work on part A, charting changes over time; and part B, charting largest change. (We later scrapped the idea of working in subteams on this part; we worked as a full team on all of it)

**Due by 1pm Wednesday**

EDA: reference kanban board: <https://github.com/NSS-Full-Time-Data-Analytics-1/healthcare-bluebook-phoenix/projects/1>

Wednesday afternoon: refine visualizations

**Due by end of day Thursday**

Analysis *(make final visualizations - some in Python, others in PowerBI/Tableau) - b*y goal (payments, utilization counts)--

**Due by Friday 1 pm (walkthrough)**

Internal presentation (TBD)

------

MVP: Set Goal analysis and presentation:

**Questions,** from Goals 2 pages from HCBB presentation (pgs 32-33)

* Which procedures *(HCPCS Codes)* had the largest change in Average payment *(using Number of Distinct Medicare Beneficiary/Per Day Services, per pg 12 of HCBB presentation*?
* Utilization?

**MVP**

* + Dashboard showing cost/utilization over the 3 years
  + **Chart showing procedures with largest change in avg payment**
    - OPTIONAL/Stretch: Interesting to see if we can find total # people covered under medicare each year, to show change as change in %. *See link below for annual # of beneficiaries*
  + **Chart showing largest change in utilization**
    - OPTIONAL/Stretch: Interesting to see if we can find total # people covered under medicare each year, to show change as change in %
  + Research into causes of oddities
  + Filter to specific procedure codes, providers, cities/states, etc

ETL Process: get all data, process to get it in usable form. (Extract/Transform/Load)

TOOL: Python

* 3 years of datasets (2015, 2016, 2017)
  + Hospital data… wait til later?
* Output options:
  + 1 dataframe, primary key, value columns for each year (adding columns per year)[merge]
    - Null values where provider not in each year
  + 1 dataframe, with year as a column (adding more rows) [concat]
    - Harder to catch where a primary key does not show up each year
  + 3 separate dataframes, pivot/manipulate separately, and export to other tool
* Decision, Output of ETL:
  + Step 1: 3 separate dataframes, pivot/manipulate separately
  + Step 2 if needed: combine all 3 into one dataframe in python

Analysis:

TOOL: Tableau/ powerBI

* Time visualizations
* Get insights

Presentation

TOOL: dashboard or powerpoint.

Stretch Goals

* Work if there’s time!!

HCBB slides: <https://drive.google.com/file/d/1o6yBhoNa08q0X_we6l4VLDuGkdmu7OUX/view?usp=sharing>

GitHub Repo:

<https://github.com/NSS-Full-Time-Data-Analytics-1/healthcare-bluebook-phoenix>

GitHub Project: <https://github.com/NSS-Full-Time-Data-Analytics-1/healthcare-bluebook-phoenix/projects/1?add_cards_query=is%3Aopen>

Yearly Medicare Enrollment Data:

<https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/CMSProgramStatistics/2017/2017_Enrollment>

Project Timeline:

* 5.5 days to work

------------------------------------------------------------

Calendar:

F

* HealthCare BlueBook project
* =============================
* Team/Instructor checkpoints

**Week 12 (6/1 – 6/5)**

M

* HealthCare BlueBook project

T

* HealthCare BlueBook project
* =============================
* Team/Instructor checkpoints

W

- HealthCare BlueBook project

Th

* HealthCare BlueBook project

F

* HealthCare BlueBook project
* =======================
* presentation walkthrough

**Week 13 (6/8 – 6/12)**

M

* HealthCare BlueBook project

T

* HealthCare BlueBook presentations

**WORKING NOTES**

**Naming conventions:**

* **Columns** 
  + Kept full name, replaced spaces and “/” with underscores. Otherwise kept the full column name without further edits\
* **Dataframes**
  + Initial DataFrames by year:
    - **df\_payments\_20yy**
  + Combined DataFrame - concatenated, with separate row for each year (each NPI will have three rows):
    - **df\_payments\_combined**
* **New Column for payment type description** (i..e, Doctor Only, Facility Only, or Doctor and Facility”
  + Payment\_type
* **New DataFrame to show average payment - concatenated** (year as value in ‘year’ column)
  + df\_avg\_pmt
* **New DataFrame to show average of the** number\_of\_distinct\_medicare\_beneficiary\_per\_day\_services:
  + df\_med\_services\_day
* **New DataFrame to show average payment - pivoted** (year as column head)
  + df\_pmt\_pvt
* **Dataframe of unique beneficiaries per day pivoted by year**
  + df\_bpd\_pvt
* ~~New DataFrame for number of services pivoted by year~~
  + ~~df\_services\_pvt~~
* ~~Dataframe of beneficiaries pivoted by year~~
  + ~~df\_benefs\_pvt~~

**Cleanup To Do / Cautionary Notes:**

* Zip codes: Take left 5 only (to drop zip+4), ensure that leading zeros were not dropped
* Round the amounts to nearest (dollar? cents?) *after* finish analysis
* City names: Many are abbreviated. If use this column, look into unique values

**Notes:**

* NPI is an integer which is fine because none have leading zeros, all begin with “1”.

**Team reminders**

* Naming conventions
* Clear jupyter notebook output before pushing to github
* Thorough eda
* Documenting process
* Be mindful of scope creep
* Business rules: ie, focus on the unique beneficiary per day

Tasks:

Break out dataframes into each metric (4 total)

**Payment amount** *(*based on *Average Medicare Allowed Amount)*

* Eda
* Analysis
* Visualization

Analysis by # of beneficiaries

* Eda
* Analysis
* Visualization

Analysis by # of services

* Eda
* Analysis
* Visualization

**Analysis by # of beneficiaries/day** (based on *Number of Distinct Medicare Beneficiary/Per Day Services*)

* Eda
* Analysis
* Visualization

**Unknowns / Hcbb Questions:**

* For Goal 2 - *“Which procedures (HCPCS Code) had the largest change in Average payment?*

*Utilization?”*

* is there a preferred metric for utilization?
  + There are 3 count measures:
    - # of services
    - # of beneficiaries
    - # of unique beneficiary/per day services
  + Could do work on all 3, but if there’s one that’s most useful we can focus efforts on digging deeper there.

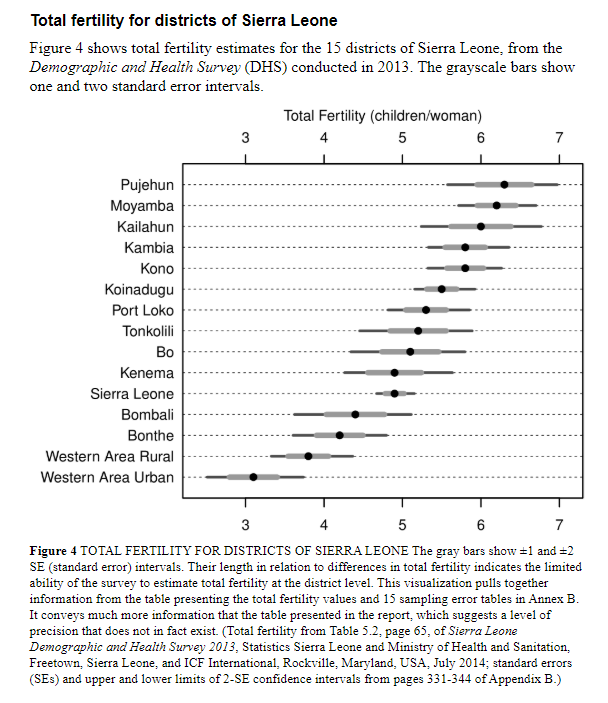
(A: Focus on the unique beneficiary per day)

**Interesting way to show change over time. Cleveland dot plot (end of dot is min/max for year**

**Viewing range of payments being collected**

**Website from Mary:** [**http://demographer.com/dsitl/08-cleveland-dot-plots/**](http://demographer.com/dsitl/08-cleveland-dot-plots/)

**Use**

****

**y procedures**

**X year**

**Pointers from Mary on presentation:**

* Restate the question we’re answering
* Use the narrative that we’ve been hired as consultants
* Talk a little bit about challenges with the data (without getting into the weeds too far)
  + Size of data
  + Creation of nulls was problematic when calculating 2017 minus 2015, e.g. provider used a code in 2015, not in 2017 causes the calculated field to be a null
  + We created a new field for Doctor Only, Facility Only, Doctor & Facility based on the criteria provided, based on the business rules provided
* For Utilization: We used the total of Doctor Only, Facility Only, and Doctor & Facility number of procedures. Using the total here gives the most clarity to see the difference between total occurrences.

*Reminder from slide deck, page 12:*

***Number of Distinct Medicare Beneficiary/Per Day Services***

*How many distinct patient/day combinations do we have for this procedure*

**Notes from presentation.**

* Clean header “where it cuts off “alo…”
* Include HCPCS Description to tables
* Research to see if hcpcs codes got changed (same description on diff hcpcs code on diff year). If so, fix this. *Possible way to check: Review on addendum Excel sheet -* ***Nicole (DONE - filtered out these codes & updated new top 10 list in slide now - showing codes that were still in use in 2017)***
* Make sure words spoken reflect what’s communicated visually.
* Change ‘soup’ to something else.
* Show specifics of what is measured on the graphs (compares 2015 to 2017)
* High level formula of what exactly went into utilization (methodology. Any important data dropped perhaps. Paying some mention of the cleaning we did of the data we’re showing. ) **-*Nicole & Diego***
* Bar charts in addition to tables.
* Map will be really cool. **:-)**
* Change **Hcpcs** to **HCPCS** in Tableau column headers **-*Nicole & Diego DONE***
* Try to remove items that are zero in 2017 - probably zero because of…. Here’s website regarding major changes to CPT codes in 2017.
  + <https://www.claconnect.com/resources/articles/ama-provides-guidance-on-2017-cpt-coding-changes-and-revised-language>
  + Add slide the summarized CPT code changes in 2017

*Note from Griffins (question 3):*

* Seemed to work well to have Tableau Public open in the browser, rather than toggling to Tableau Desktop.
* On bar chart it worked nicely to only use HCPCS code as axis label; then over the bar to see a lot more information.
* Highlighting a few examples was a really good idea (ie: sophia)
* Possibly light visuals, not necessarily a dashboard on this, for ex:
* Removing redundancies: ie: excluding office visits...no need to include bullet point since they already asked you to do that
* (“Sucker punch” ...specifically, what was it that Mary mentioned?)
* Polish & rehearse

Hexcode For HCBB:

#00A6CE

#23AFB0

#36B88F

**Team Retrospective**

**Keep Doing:**

1. Pair programming during eda
2. Sub groups focused on specific questions
3. Pivoting as needed
4. Avoiding scope creep
5. Naming conventions
6. Keeping an eye on the schedule in standups
7. Defining the MVP early
8. Detailed documentation

**Stop Doing:**

1. Don’t restrict ourselves to one tool while executing EDA. Be considerate of the *recipe*.
2. Spending too much time working in the weeds
3. “Time creep” during collaboration

**Start Doing:**

1. More intentional on creating the narrative based on data sooner than later

**Action Items:**

1. Be mindful of value of insights vs how much time being allocated
2. GitHub repo order of operations? Copy files into new personal repo. Clone. Copy over. Pull request do it after.
3. How to share Tableau? Share full cleaned up dataset (Tableau version). Zip.