EM384: Analytical Methods for Engineering Management

Lesson 6: Data Exploration and Analysis I

26 January 2023

Table of contents

1. Lesson Objectives
2. Datasets
3. Pivot Tables
4. Practical Exercise
5. Conclusion

Lesson Objectives

Lesson 6 Objectives

- Define a database and describe its components.
- Use the five-step process for data exploration to evaluate a database.
- Generate conclusions from examining relationships among variables.
- Create a ’Pivot Table’ in Excel from a dataset

Datasets

A Database/Dataset

- A database contains a dataset. A dataset is a table of information.
- In this class, we generally use two-dimensional datasets.
    - Each row is a record in the dataset.
    - Each column is a field for the records.

A Database/Dataset

- Be skeptical of data, and ask:
  - How are fields defined?
  - What types of data are represented? Nominal, Ordinal, binary, etc.
  - What units are the data in?
- 5 Steps to Using Data:
  - Understand the Data
  - Organize and Subset
  - Examine Individual Variables
  - Calculate Summary Measures for the Individual Variables
  - Examine Relationships among the Variables

The Five Steps to Using Data

The five Steps to Using Data:

- Understand the Data
- Organize and Subset
- Examine Individual Variables
- Calculate Summary Measures for the Individual Variables
- Examine Relationships among the Variables

Organize and Subset

Two essential tools: Sort and Filter.

On the Home ribbon in the Editing group and the Data Ribbon in the Sort and Filter group

Home ▸ Editing ▸ Sort & Filter ▸ Custom Sort opens the Sort window

(Sort by more than one criterion using Add Level)

Home ▸ Editing ▸ Sort & Filter ▸ Filter

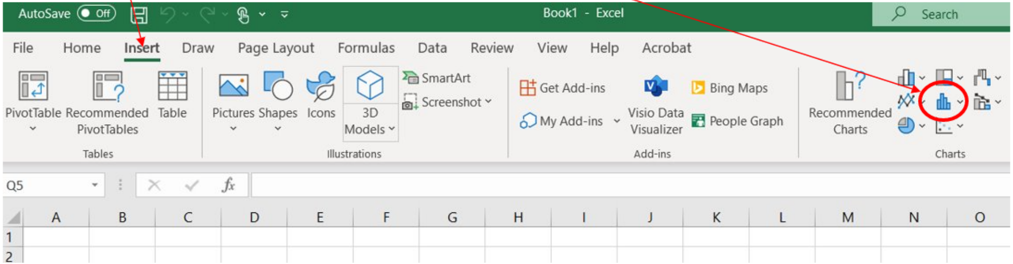
Filtering allows us to probe a large database and extract what interests us.

Warning!: Using 'average' (for example) on a filtered list will include the records that have been filtered out! Use SUBTOTAL

Examine Individual Variables

- For numerical variables, we typically want to know the range of records from lowest to highest, and areas where most outcomes lie.
- A common way to summarize a set of numerical values is the histogram, although Excel provides eight choices.
- Excel provides numerous functions useful for investigating individual variables.
- Some can summarize the values of numerical variables; others can be used to identify or count specific variables, both numerical and categorical.

- Histogram
  - Insert
    - Charts
      - Histogram



Calculate Summary Measures for Individual Variables

The most common summary measure of a numerical value is average or mean.

Calculate using the AVERAGE function in Excel, for example:

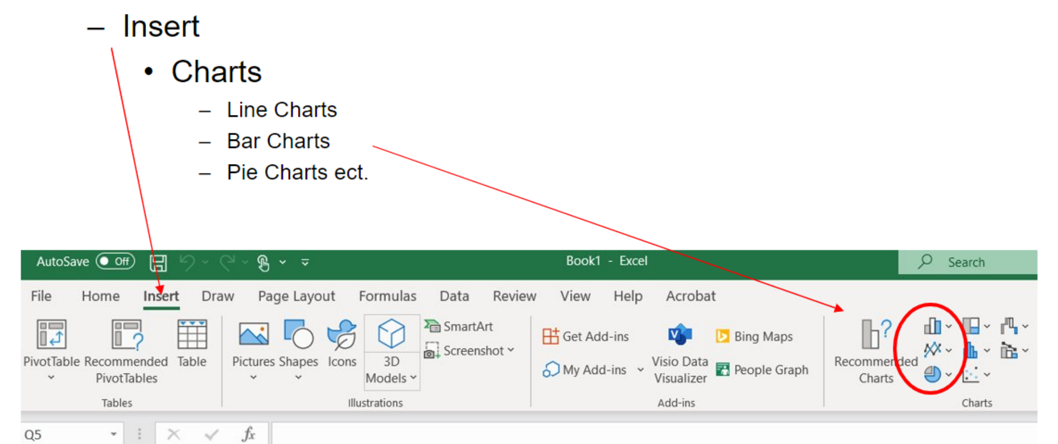
- AVERAGE(C2:C2918) = 28.97
  *Crtl-shift-down, etc!

Other useful summary measures are median, minimum, maximum, and standard deviation.

Examine Relationships

In many cases relationships among variables are more important in analysis than the properties of one variable.

Graphical methods can track relationships.



- Insert
  - Charts
    - Line Charts
    - Bar Charts
    - Pie Charts ect.

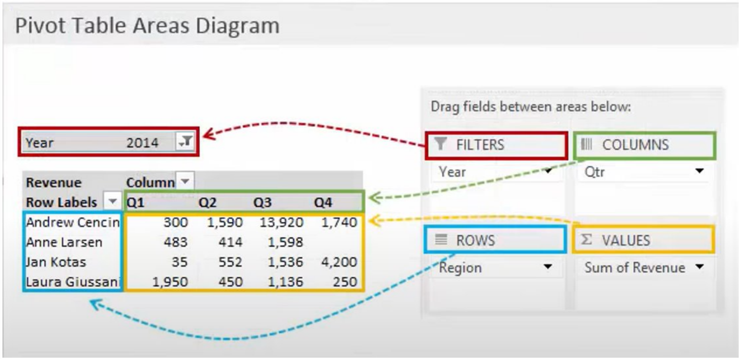
Pivot Tables

Pivot Tables

A pivot table is a table that summarizes data from another table and is made by applying an operation such as sorting, averaging, or summing to data in the first table, typically including grouping of the data.

...A tool to easily and quickly organize and present your data.

Pivot Table Areas



Pivot Table Areas Diagram

Year 2014

Revenue	Column			
Row Labels	Q1	Q2	Q3	Q4
Andrew Cencin	300	1,590	13,920	1,740
Anne Larsen	483	414	1,598	
Jan Kotas	35	552	1,536	4,200
Laura Giussani	1,950	450	1,136	250

Drag fields between areas below:

FILTERS	COLUMNS
Year	Qtr

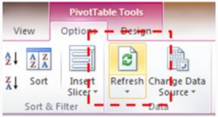
ROWS	VALUES
Region	Sum of Revenue

- Know the four options: “Filters, Rows, Columns, Values”
- You can add the same ‘field’ to multiple places
- Best technique: Trial and Error between the four options!

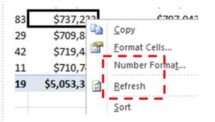
Pivot Tables - Refreshing Data

Important note about Pivot Tables: If the data values change, the table does NOT update automatically. You need to refresh the table. There are 2 ways to do this:

- From the Ribbon, under Pivot Table Tools

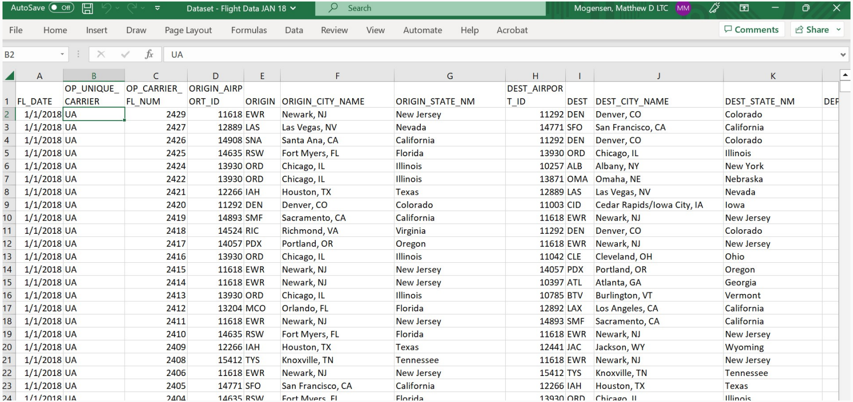


- Right Click on a call in the pivot table and choose Refresh



Pivot Table Practice

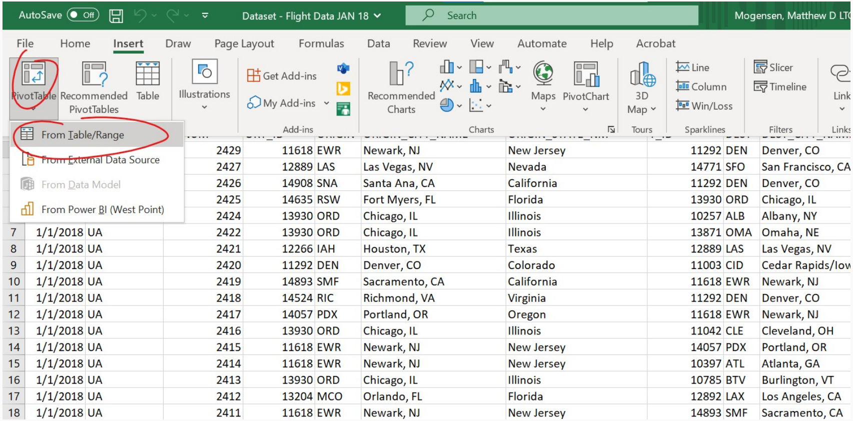
Download and open the ”Dataset - Flight Data JAN18” file from on Teams. Take a moment to examine the fields and records. How many are there of each?



	A	B	C	D	E	F	G	H	I	J	K	
1	FL_DATE	OP_UNIQUE_CARRIER	OP_CARRIER_FL_NUM	ORIGIN_AIRPORT_ID	ORIGIN	ORIGIN_CITY_NAME	ORIGIN_STATE_NM	DEST_AIRPORT_ID	DEST	DEST_CITY_NAME	DEST_STATE_NM	DEP
2	1/1/2018	UA	2429	11618	EWR	Newark, NJ	New Jersey	11292	DEN	Denver, CO	Colorado	
3	1/1/2018	UA	2427	12889	LAS	Las Vegas, NV	Nevada	14771	SFO	San Francisco, CA	California	
4	1/1/2018	UA	2426	14908	SNA	Santa Ana, CA	California	11292	DEN	Denver, CO	Colorado	
5	1/1/2018	UA	2425	14635	RSW	Fort Myers, FL	Florida	13930	ORD	Chicago, IL	Illinois	
6	1/1/2018	UA	2424	13930	ORD	Chicago, IL	Illinois	10257	ALB	Albany, NY	New York	
7	1/1/2018	UA	2422	13930	ORD	Chicago, IL	Illinois	13871	OMA	Omaha, NE	Nebraska	
8	1/1/2018	UA	2421	12266	IAH	Houston, TX	Texas	12889	LAS	Las Vegas, NV	Nevada	
9	1/1/2018	UA	2420	11292	DEN	Denver, CO	Colorado	11003	CID	Cedar Rapids/Iowa City, IA	Iowa	
10	1/1/2018	UA	2419	14893	SMF	Sacramento, CA	California	11618	EWR	Newark, NJ	New Jersey	
11	1/1/2018	UA	2418	14524	RIC	Richmond, VA	Virginia	11292	DEN	Denver, CO	Colorado	
12	1/1/2018	UA	2417	14057	PDX	Portland, OR	Oregon	11618	EWR	Newark, NJ	New Jersey	
13	1/1/2018	UA	2416	13930	ORD	Chicago, IL	Illinois	11042	CLE	Cleveland, OH	Ohio	
14	1/1/2018	UA	2415	11618	EWR	Newark, NJ	New Jersey	14057	PDX	Portland, OR	Oregon	
15	1/1/2018	UA	2414	11618	EWR	Newark, NJ	New Jersey	10397	ATL	Atlanta, GA	Georgia	
16	1/1/2018	UA	2413	13930	ORD	Chicago, IL	Illinois	10785	BTV	Burlington, VT	Vermont	
17	1/1/2018	UA	2412	13204	MCO	Orlando, FL	Florida	12892	LAX	Los Angeles, CA	California	
18	1/1/2018	UA	2411	11618	EWR	Newark, NJ	New Jersey	14893	SMF	Sacramento, CA	California	
19	1/1/2018	UA	2410	14635	RSW	Fort Myers, FL	Florida	11618	EWR	Newark, NJ	New Jersey	
20	1/1/2018	UA	2409	12266	IAH	Houston, TX	Texas	12441	JAC	Jackson, WY	Wyoming	
21	1/1/2018	UA	2408	15412	TYS	Knoxville, TN	Tennessee	11618	EWR	Newark, NJ	New Jersey	
22	1/1/2018	UA	2406	11618	EWR	Newark, NJ	New Jersey	15412	TYS	Knoxville, TN	Tennessee	
23	1/1/2018	UA	2405	14771	SFO	San Francisco, CA	California	12266	IAH	Houston, TX	Texas	
24	1/1/2018	UA	2404	14635	RSW	Fort Myers, FL	Florida	13930	ORD	Chicago, IL	Illinois	

Pivot Table Practice

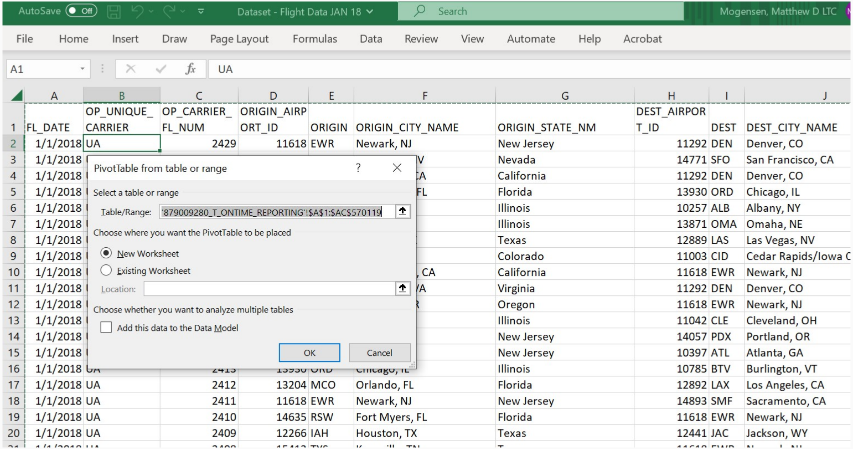
Select **insert** then **pivot table**, then **From Table/Range**.



			2429	11618	EWR	Newark, NJ	New Jersey	11292	DEN	Denver, CO
			2427	12889	LAS	Las Vegas, NV	Nevada	14771	SFO	San Francisco, CA
			2426	14908	SNA	Santa Ana, CA	California	11292	DEN	Denver, CO
			2425	14635	RSW	Fort Myers, FL	Florida	13930	ORD	Chicago, IL
			2424	13930	ORD	Chicago, IL	Illinois	10257	ALB	Albany, NY
7	1/1/2018	UA	2422	13930	ORD	Chicago, IL	Illinois	13871	OMA	Omaha, NE
8	1/1/2018	UA	2421	12266	IAH	Houston, TX	Texas	12889	LAS	Las Vegas, NV
9	1/1/2018	UA	2420	11292	DEN	Denver, CO	Colorado	11003	CID	Cedar Rapids/Iow
10	1/1/2018	UA	2419	14893	SMF	Sacramento, CA	California	11618	EWR	Newark, NJ
11	1/1/2018	UA	2418	14524	RIC	Richmond, VA	Virginia	11292	DEN	Denver, CO
12	1/1/2018	UA	2417	14057	PDX	Portland, OR	Oregon	11618	EWR	Newark, NJ
13	1/1/2018	UA	2416	13930	ORD	Chicago, IL	Illinois	11042	CLE	Cleveland, OH
14	1/1/2018	UA	2415	11618	EWR	Newark, NJ	New Jersey	14057	PDX	Portland, OR
15	1/1/2018	UA	2414	11618	EWR	Newark, NJ	New Jersey	10397	ATL	Atlanta, GA
16	1/1/2018	UA	2413	13930	ORD	Chicago, IL	Illinois	10785	BTV	Burlington, VT
17	1/1/2018	UA	2412	13204	MCO	Orlando, FL	Florida	12892	LAX	Los Angeles, CA
18	1/1/2018	UA	2411	11618	EWR	Newark, NJ	New Jersey	14893	SMF	Sacramento, CA

Pivot Table Practice

The Table/Range may autofill. Otherwise select the range for your dataset, then select **New Worksheet**, and press **OK**.



PivotTable from table or range

Select a table or range

Table/Range: '879009280_T_ONTIME_REPORTING'!$A$1:$AC$570119

Choose where you want the PivotTable to be placed

● New Worksheet

○ Existing Worksheet

Location:

Choose whether you want to analyze multiple tables

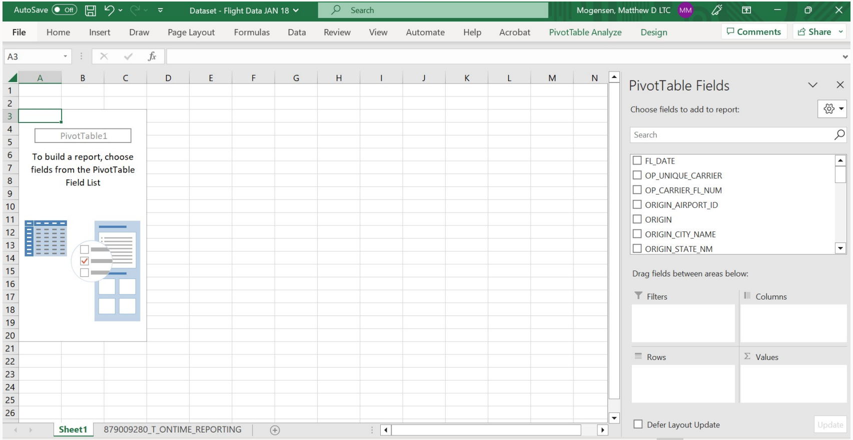
□ Add this data to the Data Model

OK Cancel

	A	B	C	D	E	F	G	H	I	J
1	FL_DATE	OP_UNIQUE_CARRIER	OP_CARRIER_FL_NUM	ORIGIN_AIRPORT_ID	ORIGIN	ORIGIN_CITY_NAME	ORIGIN_STATE_NM	DEST_AIRPORT_ID	DEST	DEST_CITY_NAME
2	1/1/2018	UA	2429	11618	EWR	Newark, NJ	New Jersey	11292	DEN	Denver, CO
3	1/1/2018					V	Nevada	14771	SFO	San Francisco, CA
4	1/1/2018					A	California	11292	DEN	Denver, CO
5	1/1/2018					FL	Florida	13930	ORD	Chicago, IL
6	1/1/2018						Illinois	10257	ALB	Albany, NY
7	1/1/2018						Illinois	13871	OMA	Omaha, NE
8	1/1/2018						Texas	12889	LAS	Las Vegas, NV
9	1/1/2018						Colorado	11003	CID	Cedar Rapids/Iowa C
10	1/1/2018					, CA	California	11618	EWR	Newark, NJ
11	1/1/2018					A	Virginia	11292	DEN	Denver, CO
12	1/1/2018					R	Oregon	11618	EWR	Newark, NJ
13	1/1/2018						Illinois	11042	CLE	Cleveland, OH
14	1/1/2018						New Jersey	14057	PDX	Portland, OR
15	1/1/2018						New Jersey	10397	ATL	Atlanta, GA
16	1/1/2018	UA	2413	13930	ORD	Chicago, IL	Illinois	10785	BTV	Burlington, VT
17	1/1/2018	UA	2412	13204	MCO	Orlando, FL	Florida	12892	LAX	Los Angeles, CA
18	1/1/2018	UA	2411	11618	EWR	Newark, NJ	New Jersey	14893	SMF	Sacramento, CA
19	1/1/2018	UA	2410	14635	RSW	Fort Myers, FL	Florida	11618	EWR	Newark, NJ
20	1/1/2018	UA	2409	12266	IAH	Houston, TX	Texas	12441	JAC	Jackson, WY

Pivot Table Practice

A new sheet should be created with the pivot table fields on the right side. If the pivot table fields are not there, click on your pivot table to the left to open the fields.

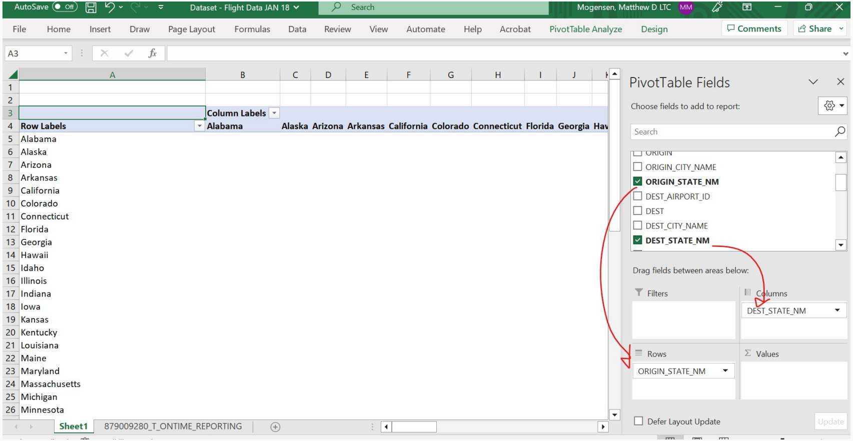


To build a report, choose fields from the PivotTable Field List

- ☐ FL_DATE
- ☐ OP_UNIQUE_CARRIER
- ☐ OP_CARRIER_FL_NUM
- ☐ ORIGIN_AIRPORT_ID
- ☐ ORIGIN
- ☐ ORIGIN_CITY_NAME
- ☐ ORIGIN_STATE_NM

Pivot Table Practice

Let’s attempt to create a pivot table that will show us the number of flights between states. Drag the **ORIGIN_STATE_NM** field to **Rows** and the **DEST_STATE_NM** to **Columns**.



	Column Labels									
Row Labels	Alabama	Alaska	Arizona	Arkansas	California	Colorado	Connecticut	Florida	Georgia	Hav
Alabama										
Alaska										
Arizona										
Arkansas										
California										
Colorado										
Connecticut										
Florida										
Georgia										
Hawaii										
Idaho										
Illinois										
Indiana										
Iowa										
Kansas										
Kentucky										
Louisiana										
Maine										
Maryland										
Massachusetts										
Michigan										
Minnesota										

PivotTable Fields

Choose fields to add to report:

Search

☐ ORIGIN
☐ ORIGIN_CITY_NAME
☑ **ORIGIN_STATE_NM**
☐ DEST_AIRPORT_ID
☐ DEST
☐ DEST_CITY_NAME
☑ **DEST_STATE_NM**

Drag fields between areas below:

Filters

Columns: DEST_STATE_NM

Rows: ORIGIN_STATE_NM

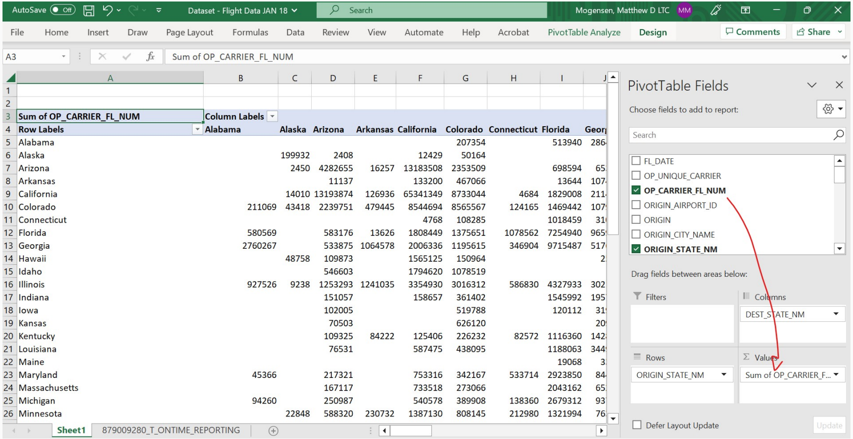
Values

☐ Defer Layout Update

Update

Pivot Table Practice

Now drag the **OP CARRIER FLIGHT NUM** field to **Values**. The default is sum. Right now this pivot table doesn't tell us much because it's summing the flight numbers.



Sum of OP_CARRIER_FL_NUM

Sum of OP_CARRIER_FL_NUM	Column Labels								
Row Labels	Alabama	Alaska	Arizona	Arkansas	California	Colorado	Connecticut	Florida	Geor
Alabama						207354		513940	286
Alaska		199932	2408		12429	50164			
Arizona		2450	4282655	16257	13183508	2353509		698594	65
Arkansas			11137		133200	467066		13644	107
California		14010	13193874	126936	65341349	8733044	4684	1829008	211
Colorado	211069	43418	2239751	479445	8544694	8565567	124165	1469442	107
Connecticut					4768	108285		1018459	31
Florida	580569		583176	13626	1808449	1375651	1078562	7254940	965
Georgia	2760267		533875	1064578	2006336	1195615	346904	9715487	517
Hawaii		48758	109873		1565125	150964			2
Idaho			546603		1794620	1078519			
Illinois	927526	9238	1253293	1241035	3354930	3016312	586830	4327933	302
Indiana			151057		158657	361402		1545992	195
Iowa			102005			519788		120112	31
Kansas			70503			626120			20
Kentucky			109325	84222	125406	226232	82572	1116360	142
Louisiana			76531		587475	438095		1188063	344
Maine								19068	3
Maryland	45366		217321		753316	342167	533714	2923850	84
Massachusetts			167117		733518	273066		2043162	65
Michigan	94260		250987		540578	389908	138360	2679312	93
Minnesota		22848	588320	230732	1387130	808145	212980	1321994	76

PivotTable Fields

Choose fields to add to report:

Search

- □ FL_DATE
- □ OP_UNIQUE_CARRIER
- ☑ **OP_CARRIER_FL_NUM**
- □ ORIGIN_AIRPORT_ID
- □ ORIGIN
- □ ORIGIN_CITY_NAME
- ☑ **ORIGIN_STATE_NM**

Drag fields between areas below:

Filters

Columns: DEST_STATE_NM

Rows: ORIGIN_STATE_NM

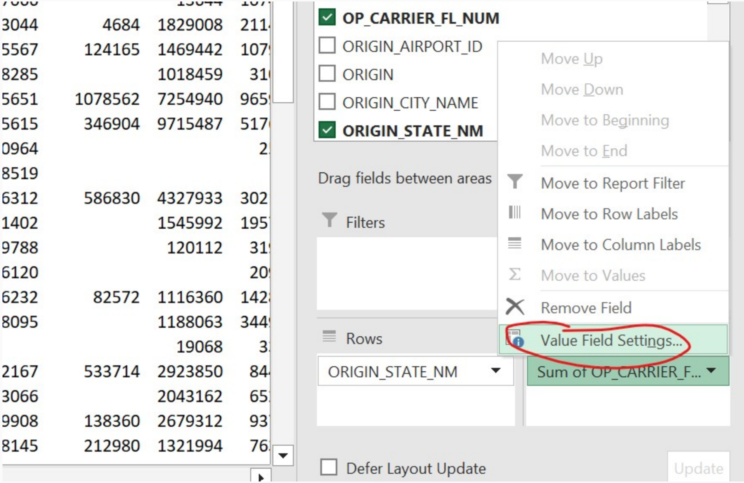
Values: Sum of OP_CARRIER_F...

□ Defer Layout Update

Update

Pivot Table Practice

Click on **OP CARRIER FLIGHT NUM** in the **Values** field.



3044	4684	1829008	211
5567	124165	1469442	107
8285		1018459	31
5651	1078562	7254940	965
5615	346904	9715487	517
0964			2
8519			
6312	586830	4327933	302
1402		1545992	195
9788		120112	31
6120			20
6232	82572	1116360	142
8095		1188063	344
		19068	3
2167	533714	2923850	84
3066		2043162	65
9908	138360	2679312	93
8145	212980	1321994	76

☑ **OP_CARRIER_FL_NUM**
□ ORIGIN_AIRPORT_ID
□ ORIGIN
□ ORIGIN_CITY_NAME
☑ **ORIGIN_STATE_NM**

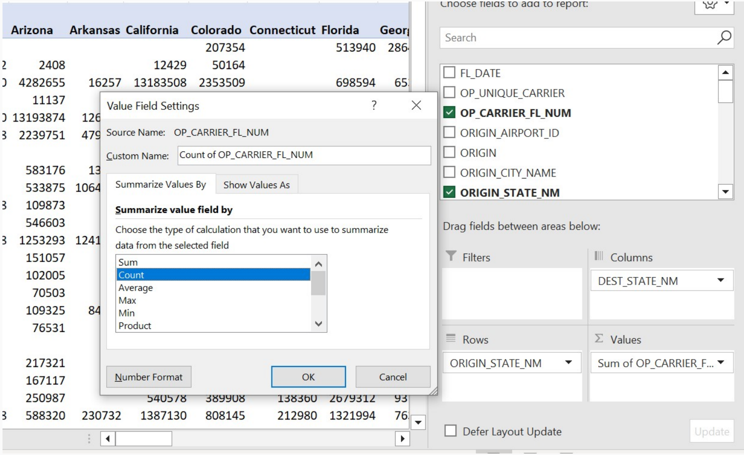
Rows: ORIGIN_STATE_NM

Values: Sum of OP_CARRIER_F...

□ Defer Layout Update

Pivot Table Practice

Select **Count**. Now it only counts the number of values instead of summing.



	Arizona	Arkansas	California	Colorado	Connecticut	Florida	Geor
				207354		513940	286
2	2408		12429	50164			
)	4282655	16257	13183508	2353509		698594	65
	11137						
)	13193874	126					
3	2239751	479					
	583176	13					
	533875	1064					
3	109873						
	546603						
3	1253293	1241					
	151057						
	102005						
	70503						
	109325	84					
	76531						
	217321						
	167117						
	250987		540578	389908	138360	2679312	93
3	588320	230732	1387130	808145	212980	1321994	76

Value Field Settings

Source Name: OP_CARRIER_FL_NUM

Custom Name: Count of OP_CARRIER_FL_NUM

Summarize value field by

Choose the type of calculation that you want to use to summarize data from the selected field

- Sum
- Count
- Average
- Max
- Min
- Product

Choose fields to add to report:

- ☐ FL_DATE
- ☐ OP_UNIQUE_CARRIER
- ☑ **OP_CARRIER_FL_NUM**
- ☐ ORIGIN_AIRPORT_ID
- ☐ ORIGIN
- ☐ ORIGIN_CITY_NAME
- ☑ **ORIGIN_STATE_NM**

Drag fields between areas below:

Filters

Columns: DEST_STATE_NM

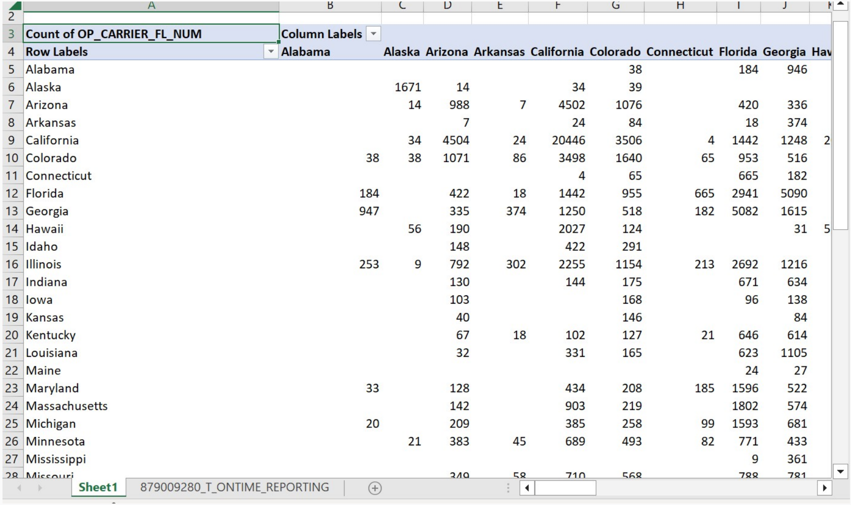
Rows: ORIGIN_STATE_NM

Values: Sum of OP_CARRIER_F...

☐ Defer Layout Update

Pivot Table Practice

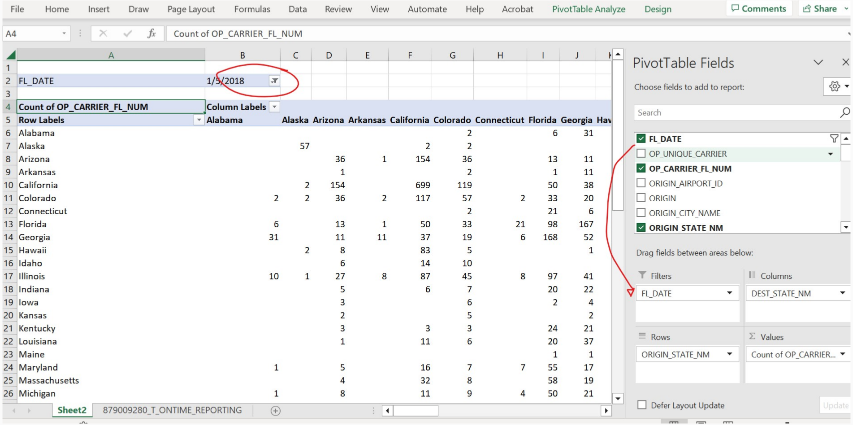
We have now completed our pivot table. What if we want to filter by date?



	A	B	C	D	E	F	G	H	I	J	K
2											
3	Count of OP_CARRIER_FL_NUM	Column Labels ▾									
4	Row Labels ▾	Alabama	Alaska	Arizona	Arkansas	California	Colorado	Connecticut	Florida	Georgia	Hav
5	Alabama						38		184	946	
6	Alaska		1671	14		34	39				
7	Arizona		14	988	7	4502	1076		420	336	
8	Arkansas			7		24	84		18	374	
9	California		34	4504	24	20446	3506	4	1442	1248	2
10	Colorado	38	38	1071	86	3498	1640	65	953	516	
11	Connecticut					4	65		665	182	
12	Florida	184		422	18	1442	955	665	2941	5090	
13	Georgia	947		335	374	1250	518	182	5082	1615	
14	Hawaii		56	190		2027	124			31	5
15	Idaho			148		422	291				
16	Illinois	253	9	792	302	2255	1154	213	2692	1216	
17	Indiana			130		144	175		671	634	
18	Iowa			103			168		96	138	
19	Kansas			40			146			84	
20	Kentucky			67	18	102	127	21	646	614	
21	Louisiana			32		331	165		623	1105	
22	Maine								24	27	
23	Maryland	33		128		434	208	185	1596	522	
24	Massachusetts			142		903	219		1802	574	
25	Michigan	20		209		385	258	99	1593	681	
26	Minnesota		21	383	45	689	493	82	771	433	
27	Mississippi								9	361	
28	Missouri			349	58	710	568		788	781	

Pivot Table Practice

Let’s display only the flights for 5 January 2018. Drag **FL_DATE** down to **Filter**. The click on the filter box on the left and select only the date we want.



Count of OP_CARRIER_FL_NUM

	A	B	C	D	E	F	G	H	I	J	Hav
1											
2	FL_DATE	1/5/2018									
3											
4	Count of OP_CARRIER_FL_NUM	Column Labels									
5	Row Labels	Alabama	Alaska	Arizona	Arkansas	California	Colorado	Connecticut	Florida	Georgia	Hav
6	Alabama						2		6	31	
7	Alaska		57			2	2				
8	Arizona			36	1	154	36		13	11	
9	Arkansas			1			2		1	11	
10	California		2	154		699	119		50	38	
11	Colorado	2	2	36	2	117	57	2	33	20	
12	Connecticut						2		21	6	
13	Florida	6		13	1	50	33	21	98	167	
14	Georgia	31		11	11	37	19	6	168	52	
15	Hawaii		2	8		83	5			1	
16	Idaho			6		14	10				
17	Illinois	10	1	27	8	87	45	8	97	41	
18	Indiana			5		6	7		20	22	
19	Iowa			3			6		2	4	
20	Kansas			2			5			2	
21	Kentucky			3		3	3		24	21	
22	Louisiana			1		11	6		20	37	
23	Maine								1	1	
24	Maryland	1		5		16	7	7	55	17	
25	Massachusetts			4		32	8		58	19	
26	Michigan	1		8		11	9	4	50	21	

PivotTable Fields

Choose fields to add to report:

- ☑ FL_DATE
- ☐ OP_UNIQUE_CARRIER
- ☑ OP_CARRIER_FL_NUM
- ☐ ORIGIN_AIRPORT_ID
- ☐ ORIGIN
- ☐ ORIGIN_CITY_NAME
- ☑ ORIGIN_STATE_NM

Drag fields between areas below:

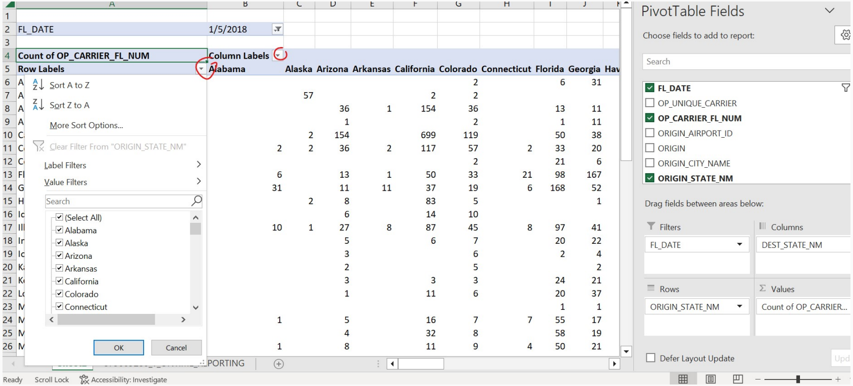
Filters	Columns
FL_DATE	DEST_STATE_NM

Rows	Values
ORIGIN_STATE_NM	Count of OP_CARRIER...

☐ Defer Layout Update

Pivot Table Practice

What if we want to only show flights between certain states? We can click the drop down buttons for rows and columns on the pivot table itself and only select the first four states.



FL_DATE	1/5/2018									
Count of OP_CARRIER_FL_NUM	Column Labels									
Row Labels	Alabama	Alaska	Arizona	Arkansas	California	Colorado	Connecticut	Florida	Georgia	Hav
A						2		6	31	
A		57			2	2				
A			36	1	154	36		13	11	
A			1			2		1	11	
C		2	154		699	119		50	38	
C	2	2	36	2	117	57	2	33	20	
C						2		21	6	
Fl	6		13	1	50	33	21	98	167	
G	31		11	11	37	19	6	168	52	
H		2	8		83	5			1	
Ic			6		14	10				
Ill	10	1	27	8	87	45	8	97	41	
Ir			5		6	7		20	22	
Ic			3			6		2	4	
K			2			5			2	
K			3		3	3		24	21	
L			1		11	6		20	37	
M								1	1	
M	1		5		16	7	7	55	17	
M			4		32	8		58	19	
M	1		8		11	9	4	50	21	

PivotTable Fields

Choose fields to add to report:

Search

- ☑ FL_DATE
- □ OP_UNIQUE_CARRIER
- ☑ OP_CARRIER_FL_NUM
- □ ORIGIN_AIRPORT_ID
- □ ORIGIN
- □ ORIGIN_CITY_NAME
- ☑ ORIGIN_STATE_NM

Drag fields between areas below:

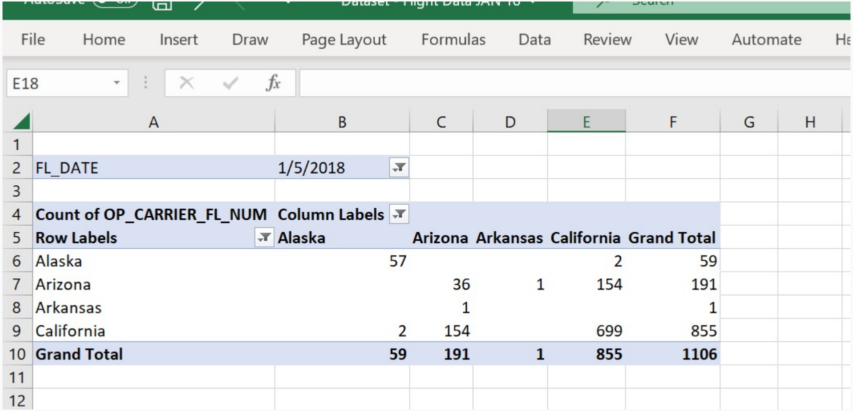
Filters	Columns
FL_DATE	DEST_STATE_NM

Rows	Values
ORIGIN_STATE_NM	Count of OP_CARRIER...

□ Defer Layout Update

Pivot Table Practice

You should end up with the following pivot table



FL_DATE 1/5/2018

Count of OP_CARRIER_FL_NUM	Column Labels				
Row Labels	Alaska	Arizona	Arkansas	California	Grand Total
Alaska	57			2	59
Arizona		36	1	154	191
Arkansas		1			1
California	2	154		699	855
Grand Total	**59**	**191**	**1**	**855**	**1106**

Practical Exercise

Conclusion

Next Class

Homework:

- Watch tutorial video on Python libraries, data structures and control structures (Will be available on Teams).

Next Lesson:

- Understand the use of Python libraries.
- Understand basic Python data structures (integer, float, boolean, string, and lists/Numpy arrays), including assignment of values and referencing.
- Understand basic Python control structures (IF,ELSE,FOR).