MOTOR VEHICLE DEATHS IN THE U.S.

An analysis of seasonality, days of the week, and age from 2005-2015

Group #3

Amanda Enstad

Barbara MacGregor

Matt Russell

Chi Tran

WHY THIS TOPIC?

- Detailed information, Well-Structured
- Significantly Large Dataset
- Easily Accessible (Kaggle)
- Trustworthy Source

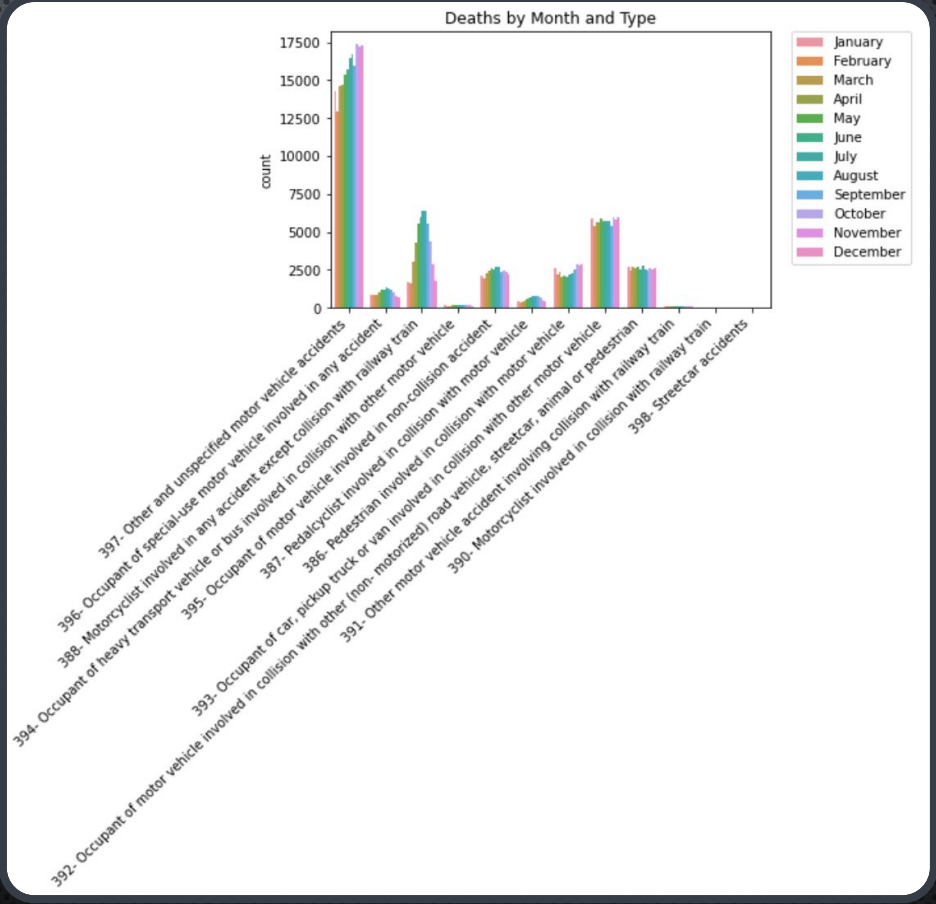
SOURCE

- Dataset: Deaths in the United States (CDC, 2017)
  - Causes of death from 2005-2015
  - Over 600 source codes (categories) in data set
    - Focused on motor vehicle deaths: 12 source codes
- CSVs for each year

- Secondary Data Source: US Census Data

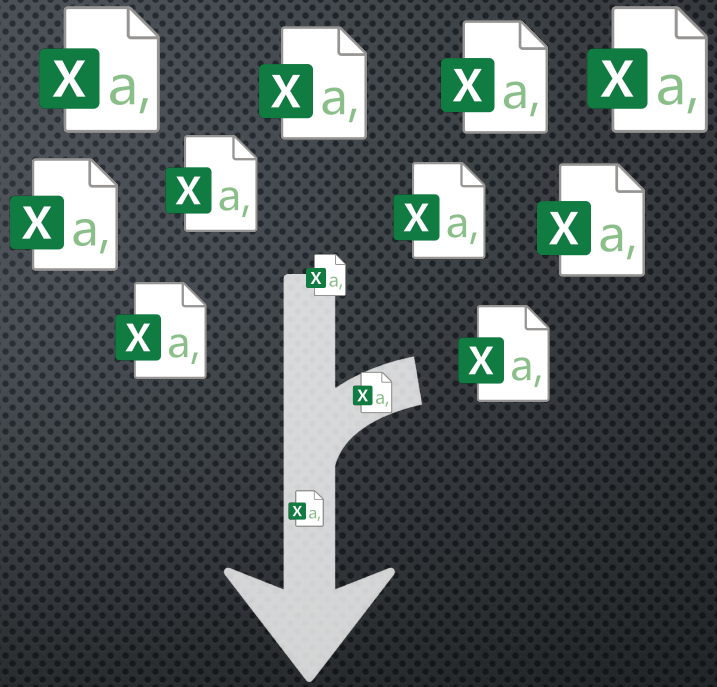
LIMITATIONS

- Lack of driving & location data
- A large percentage of motor vehicles deaths were "unspecified"
  - 45.1% of motor vehicles deaths reported were considered unspecified
- Some statistical tests/hypotheses we wanted to test were not applicable
  - Data did not have equal variances/distribution in some cases



DATA WRANGLING

STEP 1: MERGE ALL OF THE CSVS INTO A SINGLE DATAFRAME, USING ONLY THE COLUMNS WE NEED



	month_of_death	sex	detail_age	day_of_week_of_death	current_data_year	manner_of_death	358_cause_recode
207	1	M	32	7	2005	1.0	396
208	1	M	75	5	2005	1.0	387
220	1	M	68	7	2005	1.0	396
234	2	M	21	1	2005	1.0	396
235	2	M	24	1	2005	1.0	396

Step 2: Used JSON files to transform our coded dataframe into something readable

CDC definitions (JSON format)

```
manner_of_death_dict = {
      1:"Accident",
      2:"Suicide",
      3:"Homicide",
      4:"Pending investigation",
      5:"Could not determine",
      6:"Self-Inflicted",
      7:"Natural"}
#       "Blank":"Not specified"}

cause_recode_dict = {
      385:" 385- Motor vehicle accidents",
      386:" 386- Pedestrian involved in collision with motor vehicle",
      387:" 387- Pedalcyclist involved in collision with motor vehicle",
      388:" 388- Motorcyclist involved in any accident except collision with railway train",
      389:" 389- Motor vehicle accident involving collision with railway train",
      390:" 390- Motorcyclist involved in collision with railway train",
      391:" 391- Other motor vehicle accident involving collision with railway train",
      392:" 392- Occupant of motor vehicle involved in collision with other (non- motorized) road vehicle, streetcar, animal
      393:" 393- Occupant of car, pickup truck or van involved in collision with other motor vehicle",
      394:" 394- Occupant of heavy transport vehicle or bus involved in collision with other motor vehicle",
      395:" 395- Occupant of motor vehicle involved in non-collision accident",
      396:" 396- Occupant of special-use motor vehicle involved in any accident",
      397:" 397- Other and unspecified motor vehicle accidents",
      398:" 398- Streetcar accidents"}


clean_df = car_death_data.replace({"month_of_death": month_dict,
                        "day_of_week_of_death": day_of_week_dict,
                        "manner_of_death": manner_of_death_dict,
                        "358_cause_recode": cause_recode_dict})

clean_df
```

Replacing coded values in dataframe with new values from dictionaries

Step 2: Used json files to transform our coded dataframe into something readable

The resulting dataframe is clean, easy to understand & ready for analysis!

	month_of_death	sex	detail_age	day_of_week_of_death	current_data_year	manner_of_death	358_cause_recode
207	January	M	32	Saturday	2005	Accident	396- Occupant of special-use motor vehicle in...
208	January	M	75	Thursday	2005	Accident	387- Pedalcyclist involved in collision with ...
220	January	M	68	Saturday	2005	Accident	396- Occupant of special-use motor vehicle in...
234	February	M	21	Sunday	2005	Accident	396- Occupant of special-use motor vehicle in...
235	February	M	24	Sunday	2005	Accident	396- Occupant of special-use motor vehicle in...
...	...	...	...	...	...	...	...
2717184	December	M	67	Friday	2015	Accident	397- Other and unspecified motor vehicle acci...
2717278	December	F	77	Thursday	2015	Accident	397- Other and unspecified motor vehicle acci...
2717674	December	M	73	Wednesday	2015	Accident	397- Other and unspecified motor vehicle acci...
2717998	December	M	70	Thursday	2015	Accident	395- Occupant of motor vehicle involved in no...
2718169	December	F	63	Thursday	2015	Accident	386- Pedestrian involved in collision with mo...

Step 3: Check the data quality & remove any inaccurate or ambiguous data

In [5]:

```
#  check quality of data
#  list all unique values in each columns

colNames = list(clean_df.columns)
for col in colNames:
    print(col)
    print(f"{clean_df[col].unique()}")
    print("----------")
```

```
month_of_death
['January' 'February' 'March' 'April' 'June' 'May' 'July' 'August'
 'September' 'October' 'November' 'December']
----------
sex
['M' 'F']
----------
detail_age
[ 32  75  68  21  24  25  44  49  40  11  14  64  57  18   9  54  31  42
  27  43  58  46  33  22  13  61  56  16  53  30  23  26  20  83  34  38
  85  15   1  47  65  72  52  17  28  12  19  48  45  50  66  81  41  62
  59  69  39  37  63  78  80  35  73  36  90 999  82  55  51   2  60  70
  29  91  79   7  84  76   8   5  77  89  74   4  10  88  86   6   3  71
  67  92 101  94  87  93  97  96 102 100  99  95  98 106 104 103 107]
----------
day_of_week_of_death
['Saturday' 'Thursday' 'Sunday' 'Tuesday' 'Wednesday' 'Monday' 'Friday'
 'Unknown']
----------
current_data_year
[2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015]
----------
manner_of_death
['Accident' 'Natural' 'Could not determine' 'Homicide' nan
 'Pending investigation' 'Suicide']
----------
358_cause_recode
[' 396- Occupant of special-use motor vehicle involved in any accident'
 ' 387- Pedalcyclist involved in collision with motor vehicle'
```

In [6]:

```
#  Clean up
#  Remove not logical data
# ie: age of 999, day of week : Unknown, manner_of_death nan, need filter by Accident
finalDf = clean_df[clean_df["detail_age"] != 999]
finalDf = finalDf[finalDf["manner_of_death"] == "Accident"]
finalDf =  finalDf[finalDf["day_of_week_of_death"] != "Unknown"]
finalDf.head()
```


NOW, WE CAN USE OUR FINAL, CLEAN DATAFRAME TO ANSWER
OUR RESEARCH QUESTIONS!

Does the average number of vehicle deaths vary by month?

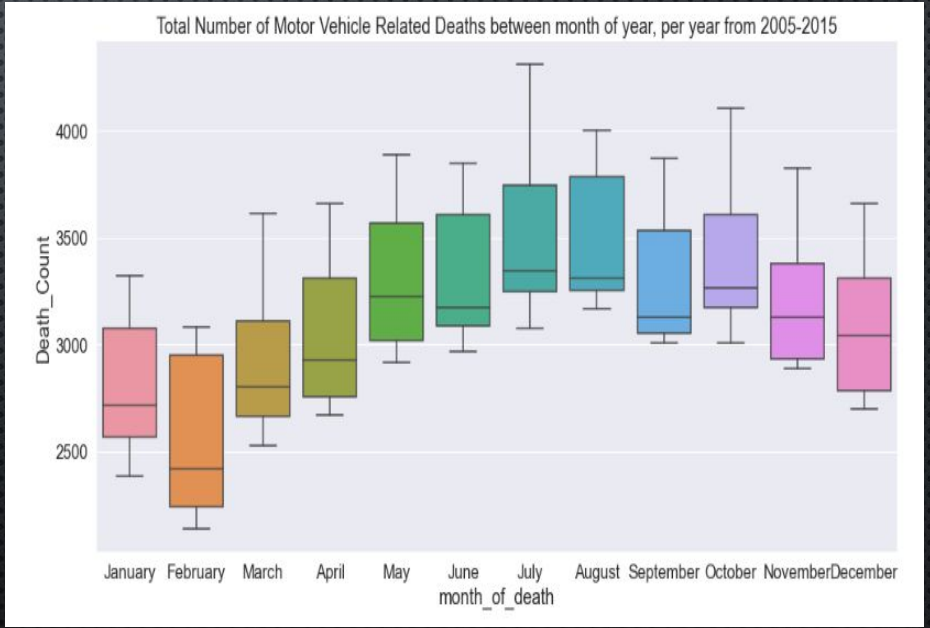
1.) Group the dataframe by month, then sum how many deaths occurred in each month per year

2.) Levene's Test to test for equal variance

3.) ANOVA to determine if there is a difference between groups in our population

4.) f-Oneway Test to compare specific groups

```
================
Levene Summary
================
    W    pval  equal_var
-----  ------  -----------
0.072   1.000  True


=============
ANOVA SUMMARY
=============

Source            ddof1    ddof2      F    p-unc    np2
--------------  -------  -------  -----  -------  -----
month_of_death       11      120  7.664    0.000  0.413

================
Feb vs. Aug
================
F_onewayResult(statistic=42.425595897501985, pvalue=2.385867742925349e-06)
================
Jul vs. Aug
================
F_onewayResult(statistic=0.00010788018826291518, pvalue=0.9918157990446276)
```



Does the average number of vehicle deaths vary by season?

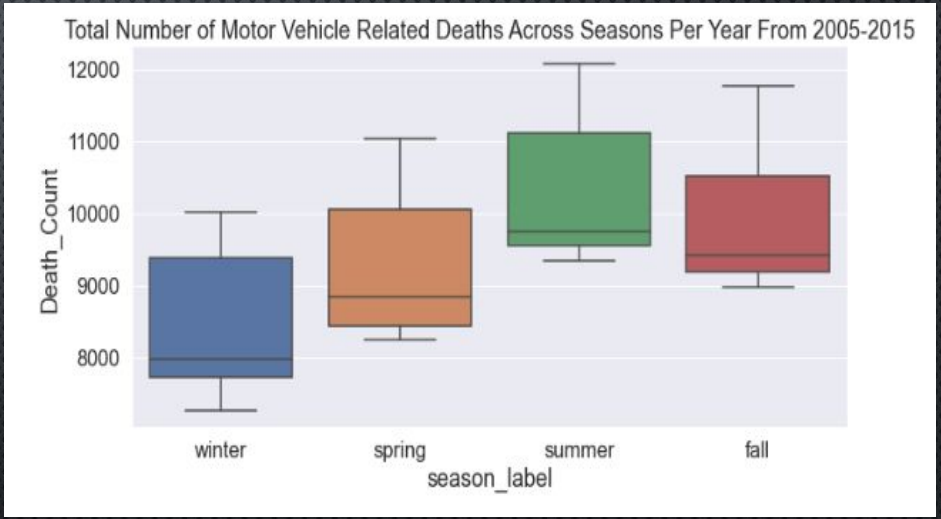
1.) Season definitions

A. Winter: Dec-Feb
B. Spring: March – May
C. Summer: June-Aug
D. Fall: Sept-Nov

2.) Group the dataframe by season, then sum how many deaths occurred in each season per year

3.) Levene's Test to test for equal variance

4.) ANOVA to determine if there is a difference between groups in our population

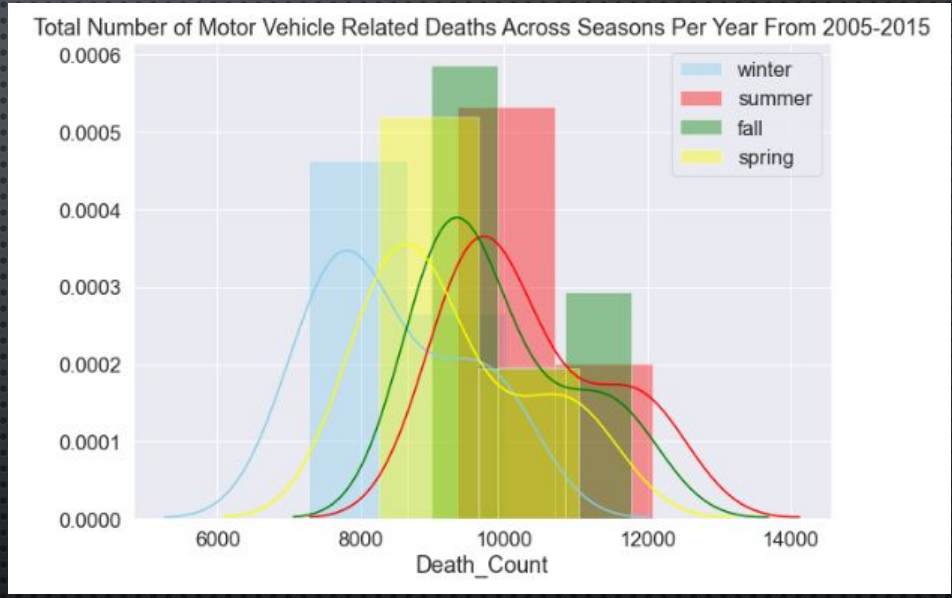


Does the average number of vehicle deaths vary by season?

1.) p-value <0.05 from ANOVA reject null hypothesis

2.) Pairwise Tukey test to compare different seasons

3.) Comparing summer and winter, Pairwise Tukey and F-Oneway have different results



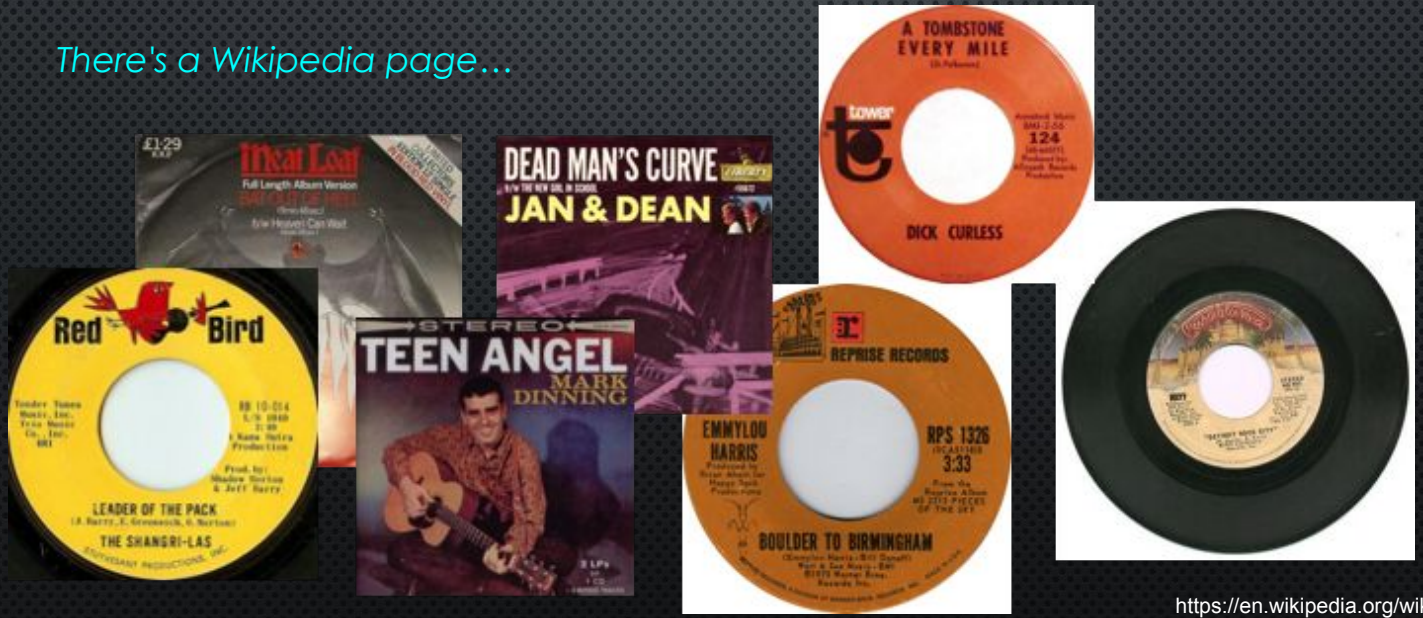

```
==============
POST HOC TESTS
==============

A       B         mean(A)    mean(B)       diff       se  tail           T    p-tukey    hedges
------  ------  ---------  ---------  ---------  -------  ---------  ------  ---------  --------
fall    spring    8441.000   9267.182   -826.182  445.349  two-sided  -1.855      0.252    -0.761
fall    summer    8441.000  10350.000  -1909.000  445.349  two-sided  -4.287      0.001    -1.758
fall    winter    8441.000   9940.455  -1499.455  445.349  two-sided  -3.367      0.005    -1.381
spring  summer    9267.182  10350.000  -1082.818  445.349  two-sided  -2.431      0.076    -0.997
spring  winter    9267.182   9940.455   -673.273  445.349  two-sided  -1.512      0.433    -0.620
summer  winter   10350.000   9940.455    409.545  445.349  two-sided   0.920      0.768     0.377

================
FOneWay : Winter vs. Summer
================

F_onewayResult(statistic=18.781165691792996, pvalue=0.00032243751192492423)
```

Is there an interaction between age and day of week for motor vehicle deaths?

Specifically:

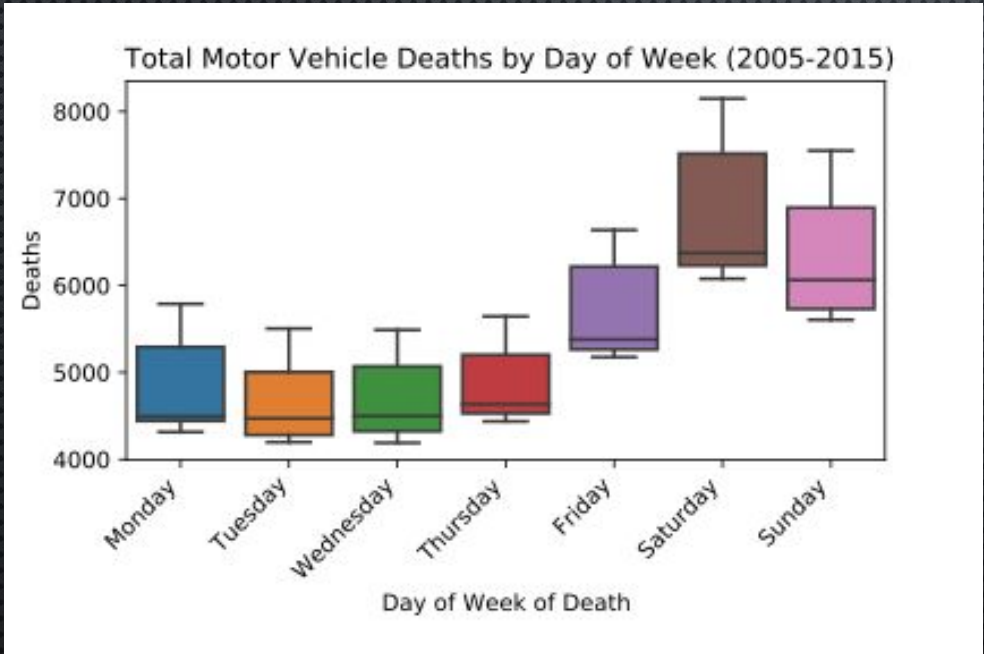
- Do younger people show a stronger tendency toward weekend fatalities?
- H_o in all cases is that there is no effect (whether age, day of week, or both)
- Motivation: Popular culture and high school experience

There's a Wikipedia page…



https://en.wikipedia.org/wiki/List_of_car_crash_songs

Motor vehicles deaths by day of week



Levene's Test

p-value = 8.27 e-01 (same for stats package and pingouin)

P >= 0.05

➢ Sample variances similar enough for ANOVA

ANOVA F-Oneway

Statistic = 7.66

P = 6.62 e-10 (stats package)

P(uncorr.) = 1.1 1e-14 (Pingouin)

➢ Reject the null hypothesis that the sample means are equal

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.levene.html

Motor vehicle deaths by age group

Levene's Test

p-value = 5.79 E-09 (Pingouin package)

P >= 0.05

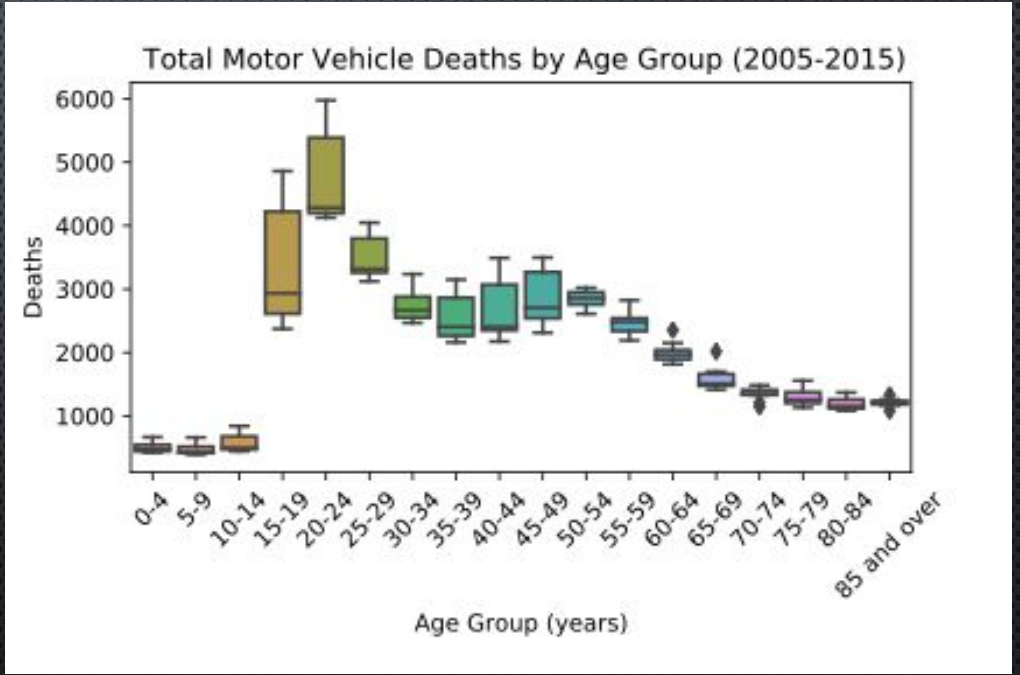
➢ Sample variances too dissimilar for ANOVA

Kruskal-Wallis H-test for independent samples

Statistic = 183.13

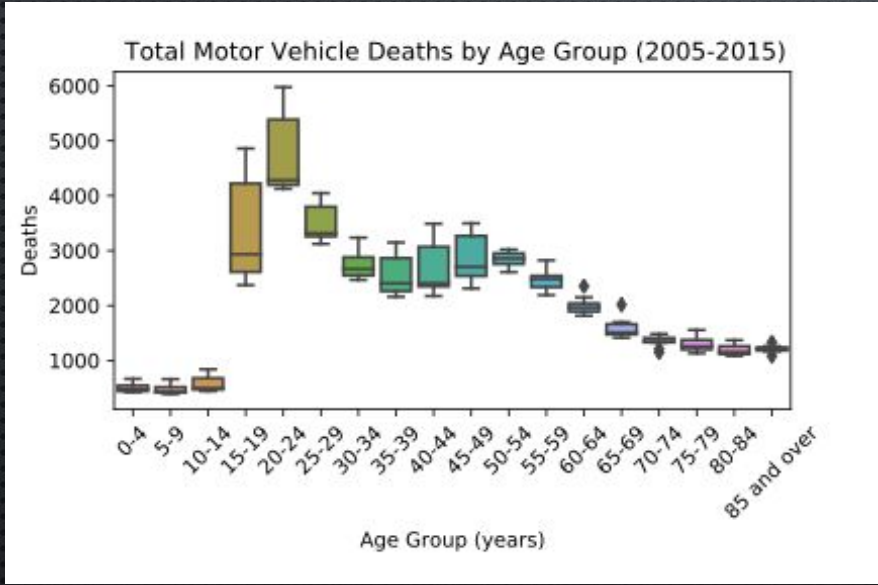
P = 6.84 e-30 (Pingouin package)

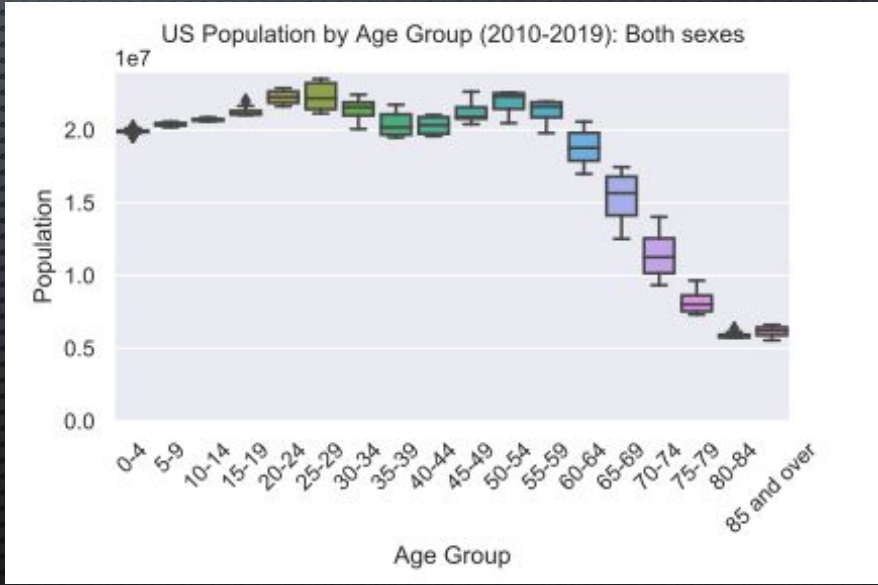
➢ Reject the null hypothesis that the sample medians are equal



https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.levene.html
https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kruskal.html

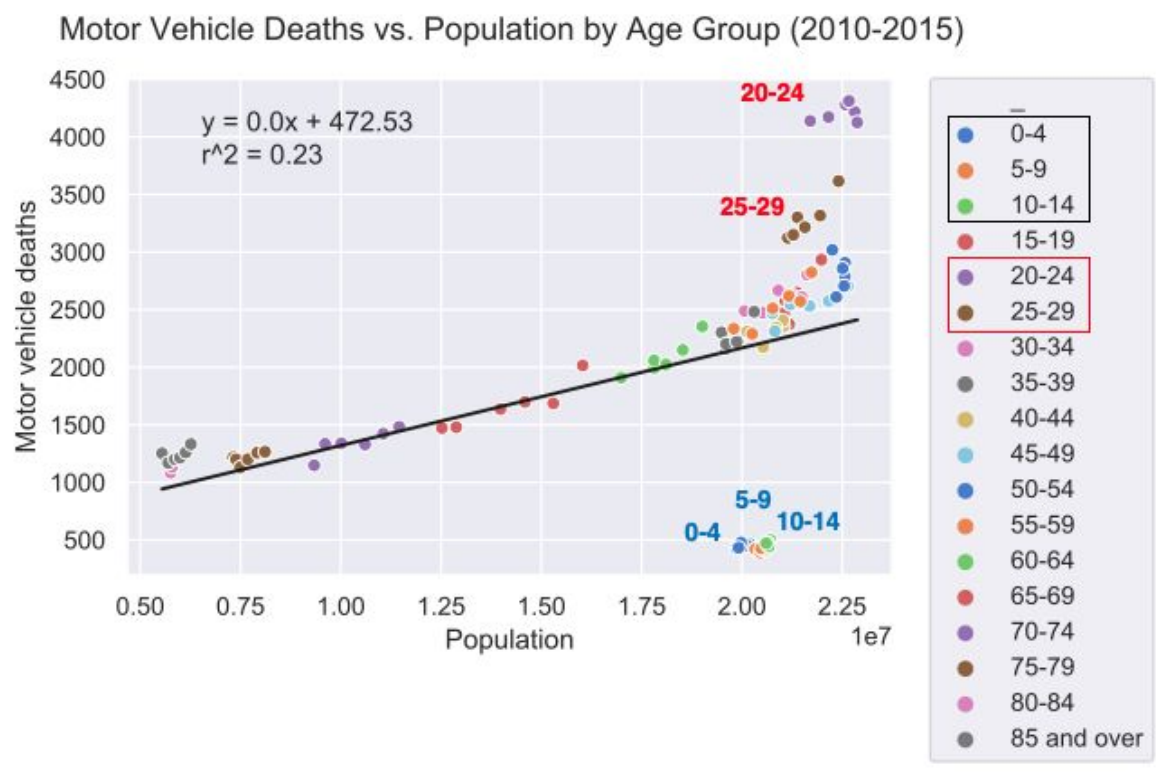
Can some age-group effects be attributed to differing cohort sizes?





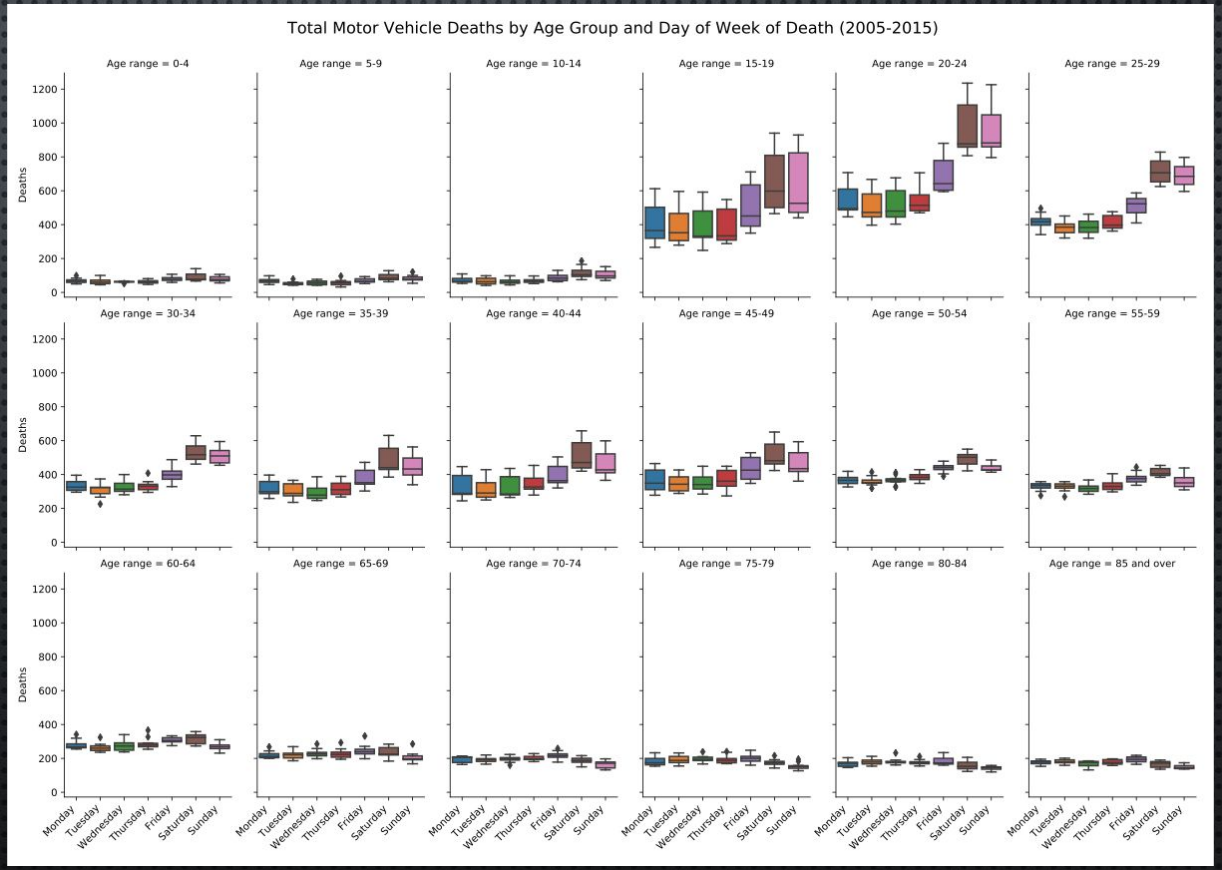
https://www.chttps://www.census.gov/data/tables/time-series/demo/popest/2010s-national-detail.html

Motor vehicle deaths by age group



Linear regression of motor vehicle deaths vs. population suggests there are other factors, particularly for children and young adults

- Age-group variances are too dissimilar for ANOVA
- Can age groups be tested one by one?
- Visual impression is that any weekend effect may dissipate with age



- Variances within age groups are suitable for ANOVA test of day-to-day differences

	W	pval	equal_var
levene	1.94635	0.085272	True
	W	pval	equal_var
levene	0.882933	0.512125	True
	W	pval	equal_var
levene	1.078084	0.384021	True
	W	pval	equal_var
levene	0.933086	0.476952	True
	W	pval	equal_var
levene	0.661903	0.680509	True
	W	pval	equal_var
levene	1.668938	0.141543	True
	W	pval	equal_var
levene	0.805239	0.569251	True
	W	pval	equal_var
levene	0.988245	0.440008	True
	W	pval	equal_var
levene	0.35972	0.901849	True
	W	pval	equal_var
levene	0.830179	0.550596	True
	W	pval	equal_var
levene	0.866909	0.523659	True
	W	pval	equal_var
levene	0.565013	0.75671	True
	W	pval	equal_var
levene	0.519195	0.791876	True
	W	pval	equal_var
levene	0.343863	0.911163	True
	W	pval	equal_var
levene	0.409808	0.870158	True
	W	pval	equal_var
levene	0.563515	0.757874	True
	W	pval	equal_var
levene	1.066215	0.391112	True
	W	pval	equal_var
levene	0.840581	0.5429	True

	p (uncorr.)
Overall	0
Age group	
0-4	0.000039
5-9	0.000002
10-14	8.33E-07
15-19	0.000006
20-24	6.80E-19
25-29	5.00E-28
30-34	2.24E-23
35-39	3.37E-12
40-44	1.43E-09
45-49	2.40E-09
50-54	1.22E-18
55-59	7.35E-11
60-64	0.000145
65-69	0.066528
70-74	0.000002
75-79	0.000216
80-84	0.000015
85 and over	4.41E-07

- One-way ANOVA showed significant differences among days for all but the 65-69 age group
- At this level of analysis, younger people are not different

Post-hoc tests showed significant differences in motor vehicle fatality counts among all pairs of weekdays and age groups, except those shown here

	Post-hoc tests corrected for multiple-comparisons												
2	Contrast	A	B	Paired	Parametric	T	dof	Tail	p-unc	p-corr	p-adjust	BF10	hedges\#NAME?
3	day_of_week_of_death	Monday	Thursday	TRUE	TRUE	-1.276	197	two-sided	0.204	0.204	fdr_bh	0.177	-0.016\
4	day_of_week_of_death	Tuesday	Wednesday	TRUE	TRUE	-1.816	197	two-sided	0.071	0.074	fdr_bh	0.398	-0.024\
5													
6	Age_group	45-49	50-54	TRUE	TRUE	0.229	76	two-sided	0.819	0.819	fdr_bh	0.129	0.023\
7	Age_group	80-84	85 and over	TRUE	TRUE	-1.039	76	two-sided	0.302	0.304	fdr_bh	0.211	-0.134\
8	Age_group	35-39	55-59	TRUE	TRUE	1.368	76	two-sided	0.175	0.178	fdr_bh	0.307	0.166\
9	Age_group	30-34	40-44	TRUE	TRUE	1.493	76	two-sided	0.14	0.142	fdr_bh	0.363	0.085\
10	Age_group	15-19	25-29	TRUE	TRUE	-1.53	76	two-sided	0.13	0.134	fdr_bh	0.383	-0.112\
11													

- Some tendency for age groups to cluster
- No evidence yet to cleanly separate weekends from weekdays
- likely next test would be to group (some) weekdays vs. weekend
- Analyses to date are easiest with equal sample sizes, so would choose e.g. Sat/Sun vs. Tue/Wed

CONCLUSIONS/ FURTHER QUESTIONS

- There are significant differences in motor vehicle deaths between:
  - Summer & Winter
  - Age
  - Days of Week

- Questions we have after our analyses:
  - What is the correlation between auto insurance rates and age of driver?
  - Future Analysis: Will motor vehicle deaths decrease with the newly enforced texting and driving laws