FITBIT ANALYSIS

VIVIAN WU

PROBLEM

- **Consumers use Fitbit to track physical activity and sleep**
- **Data is stored and accessible**
  - Users can export their own data on Fitbit's website
- **How is the tracked information used, if at all?**
  - Numbers on a spreadsheet are meaningless without interpretation

SOLUTION

- **Tracy**
  - Fitbit consumer
  - Group exercise instructor
  - Home health aide
- **Provide stats and visualizations of the data**
  - Provides insight on Tracy’s fitness & sleeping habits
  - Helps gauge what habits should be maintained or changed

THE DATA

- **Extract up to 31 days of data at a time**
- **Body, activity, sleep, food data**
- **CSV or XLSX file option**
  - Each category of data is saved in a separate sheet (within the same Excel file)

THE DATA

Each XLSX file ‘2018-10.xls’

- **“Foods”**
  - Date
  - Calories In
- **“Activities”**
  - Dates
  - Calories Burned
  - Steps
  - Distance
  - Floors
  - Minutes Sedentary
  - Minutes Lightly Active
  - Minutes Fairly Active
  - Minutes Very Active
  - Activity Calories
- **“Sleep”**
  - Start Time
  - End Time
  - Minutes Asleep
  - Minutes Awake
  - Number of Awakenings
  - Time in Bed
  - Minutes REM Sleep
  - Minutes Light Sleep
  - Minutes Deep Sleep
- **“Food Log 20181001”**
  - One for each day

THE DATA

The ideal dataset:

DataFrame object

Column names

Label index
(country code)

	Country	Popu	Percent
IT	Italy	61	0.83
ES	Spain	46	0.63
GR	Greece	11	0.15
FR	France	65	0.88
PO	Portugal	10	0.14

Data
(different type in each column)

Fitbit's activity & sleep data:

	A	B	C	D	E
1	Date	Calories Burned	Steps	Distance	Floors
2	2018-10-01	3,196	23,666	8.58	35
3	2018-10-02	2,434	13,287	4.66	11
4	2018-10-03	3,699	30,983	13.28	31
5	2018-10-04	2,471	14,381	4.85	21
6	2018-10-05	2,269	13,086	4.55	20
7	2018-10-06	3,209	31,461	16.35	25
8	2018-10-07	2,277	10,872	3.99	10
9	2018-10-08	3,278	25,599	9.6	14
10	2018-10-09	2,235	13,519	5.26	22

	A	B	C
1	Start Time	End Time	Minutes Asleep
2	2018-10-30 10:15PM	2018-10-31 6:43AM	447
3	2018-10-29 10:00PM	2018-10-30 6:49AM	467
4	2018-10-28 9:55PM	2018-10-29 6:48AM	489
5	2018-10-27 10:29PM	2018-10-28 6:40AM	444
6	2018-10-26 10:25PM	2018-10-27 7:27AM	484
7	2018-10-25 9:24PM	2018-10-26 4:59AM	412
8	2018-10-24 10:12PM	2018-10-25 6:12AM	436
9	2018-10-23 11:24PM	2018-10-24 6:33AM	383
10	2018-10-22 9:04PM	2018-10-23 5:28AM	454

THE DATA (PROBLEM)

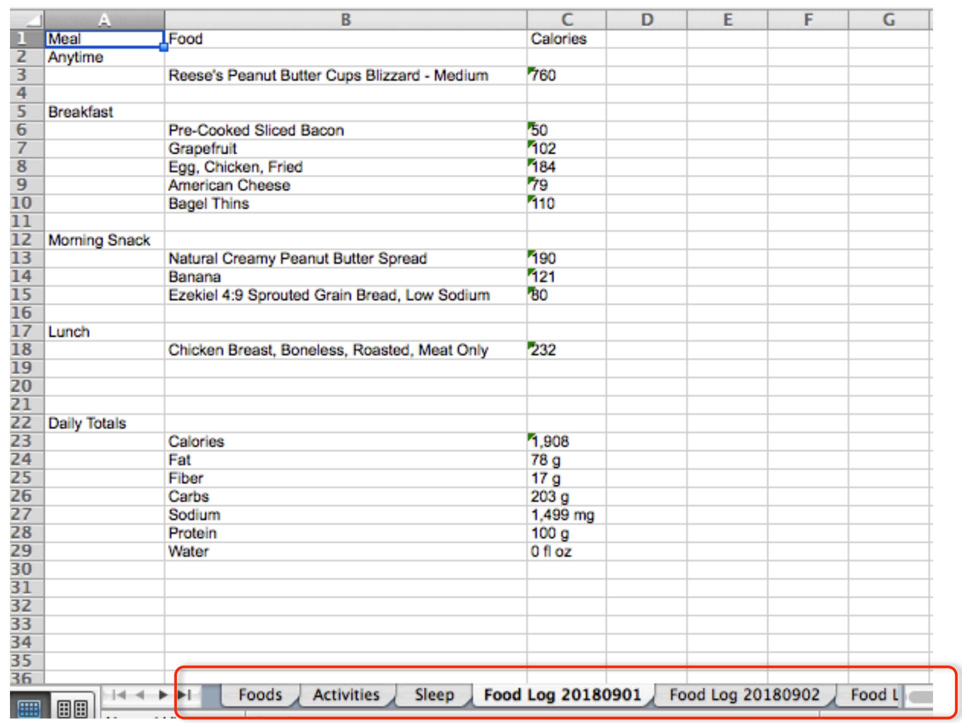
Fitbit displays food data in a “labeled” table format:

	A	B	C
1	Meal	Food	Calories
2	Breakfast		
3		English Muffin, 100% Whole Wheat	130
4		Grapefruit	102
5		Egg, Chicken, Fried	184
6		Pre-Cooked Sliced Bacon	50
7		American Cheese	79
8			
9	Morning Snack		
10		Vanilla Waffle	160
11		Banana	121
12			
13			
14			
15	Daily Totals		
16		Calories	826
17		Fat	33 g
18		Fiber	10 g
19		Carbs	101 g
20		Sodium	909 mg
21		Protein	29 g
22		Water	0 fl oz

Also includes random table with nutrition info at bottom of file

THE DATA (PROBLEM)

- **Food data unusable in the given format**
  - Need to transform
- **One sheet for each day's food log**
  - Each XLSX file contains 30+ sheets
  - Not all sheets contain the same kind of data
  - Need to determine how to to correctly extract data to the appropriate dataframe



	A	B	C	D	E	F	G
1	Meal	Food	Calories				
2	Anytime						
3		Reese's Peanut Butter Cups Blizzard - Medium	760				
4							
5	Breakfast						
6		Pre-Cooked Sliced Bacon	50				
7		Grapefruit	102				
8		Egg, Chicken, Fried	184				
9		American Cheese	79				
10		Bagel Thins	110				
11							
12	Morning Snack						
13		Natural Creamy Peanut Butter Spread	190				
14		Banana	121				
15		Ezekiel 4:9 Sprouted Grain Bread, Low Sodium	80				
16							
17	Lunch						
18		Chicken Breast, Boneless, Roasted, Meat Only	232				
19							
20							
21							
22	Daily Totals						
23		Calories	1,908				
24		Fat	78 g				
25		Fiber	17 g				
26		Carbs	203 g				
27		Sodium	1,499 mg				
28		Protein	100 g				
29		Water	0 fl oz				
30							
31							
32							
33							
34							
35							
36							

DATA WRANGLING

Food dataframe

- **Matched food to meal**
- **Added 'Date'**
  - 'Food Log 20151109'
- **Added 'Weekday'**

	A	B	C
1	Meal	Food	Calories
2	Breakfast		
3		English Muffin, 100% Whole Wheat	130
4		Grapefruit	102
5		Egg, Chicken, Fried	184
6		Pre-Cooked Sliced Bacon	50
7		American Cheese	79
8			
9	Morning Snack		
10		Vanilla Waffle	160
11		Banana	121
12			
13			
14			
15	Daily Totals		
16		Calories	826
17		Fat	33 g
18		Fiber	10 g
19		Carbs	101 g
			909 mg
			29 g
			0 fl oz

```
food.head()
```

	Meal	Food	Calories	Date	Weekday
0	Breakfast	American Cheese	61	2015-11-09	Monday
1	Breakfast	Bagel thins, Everything	110	2015-11-09	Monday
2	Breakfast	Egg, Chicken, Fried	184	2015-11-09	Monday
3	Breakfast	Ham Steak, Traditional	30	2015-11-09	Monday
4	Morning Snack	Dark Chocolate Dreams	170	2015-11-09	Monday

DATA WRANGLING

Macros dataframe

- **Pivoted original table**
- **Added 'Date' as index**
  - 'Food Log 20151109'
- **Added 'Weekday'**

	A	B	C
1	Meal	Food	Calories
2	Breakfast		
3		English Muffin, 100% Whole Wheat	130
4		Grapefruit	102
5		Egg, Chicken, Fried	184
6		Pre-Cooked Sliced Bacon	50
7		American Cheese	79
8			
9	Morning Snack		
10		Vanilla Waffle	160
11		Banana	121
12			
13			
14			
15	Daily Totals		
16		Calories	826
17		Fat	33 g
18		Fiber	10 g
19		Carbs	101 g
20		Sodium	909 mg
21		Protein	29 g
22		Water	0 fl oz

```
macros.head()
```

Date	Calories (g)	Carbs (g)	Fat (g)	Fiber (g)	Protein (g)	Sodium (mg)	Water (fl oz)	Weekday
2015-11-09	715	72	34	8	35	943	0	Monday
2015-11-11	797	74	39	4	37	1064	0	Wednesday
2015-11-12	1049	108	45	11	53	1216	0	Thursday
2015-11-30	90	20	0	1	1	2	0	Monday
2015-12-02	240	29	6	3	17	152	0	Wednesday

DATA WRANGLING

- **Replaced outliers with mean value of respective column**
- **Converted string to numeric value**
- **Remove missing values**
  - When ‘Steps’ is 0

```
activities.head()
```

Date	Calories Burned	Steps	Distance	Floors	Minutes Sedentary	Minutes Lightly Active	Minutes Fairly Active	Minutes Very Active	Activity Calories	Weekday
2015-10-21	2150.000000	14061.0	5.71	17.0	531.452381	324.060524	0.0	0.0	1588.812105	Wednesday
2015-10-22	2274.000000	13617.0	5.46	12.0	596.000000	300.000000	17.0	69.0	1344.000000	Thursday
2015-10-23	2174.000000	16530.0	6.57	20.0	639.000000	361.000000	15.0	35.0	1275.000000	Friday
2015-10-24	2161.000000	14710.0	5.88	11.0	550.000000	278.000000	36.0	52.0	1227.000000	Saturday
2015-10-25	2479.197832	5077.0	2.02	8.0	869.000000	324.060524	9.0	14.0	1588.812105	Sunday

DATA WRANGLING

- **Sleep dataframe includes nap as well**
- **Derived a 'Daily Sleep' dataframe from original 'Sleep' dataframe**

```
sleep.head()
```

	Start Time	End Time	Minutes Asleep	Minutes Awake	Number of Awakenings	Time in Bed	Date
0	2015-10-22 00:00:00	2015-10-22 05:07:00	292	15	1	307	2015-10-21
1	2015-10-22 21:29:00	2015-10-23 04:17:00	401	7	1	408	2015-10-22
2	2015-10-23 21:47:00	2015-10-24 06:43:00	514	22	2	536	2015-10-23
3	2015-10-24 23:24:00	2015-10-25 07:16:00					
4	2015-10-24 14:40:00	2015-10-24 16:05:00					

```
daily_sleep.head()
```

Date	Minutes Asleep	Minutes Awake	Number of Awakenings	Time in Bed	Weekday
2015-10-21	459.233849	15.0	1.0	496.392175	Wednesday
2015-10-22	401.000000	7.0	1.0	408.000000	Thursday
2015-10-23	514.000000	22.0	2.0	536.000000	Friday
2015-10-24	539.000000	18.0	1.0	557.000000	Saturday
2015-10-26	532.000000	33.0	2.0	565.000000	Monday

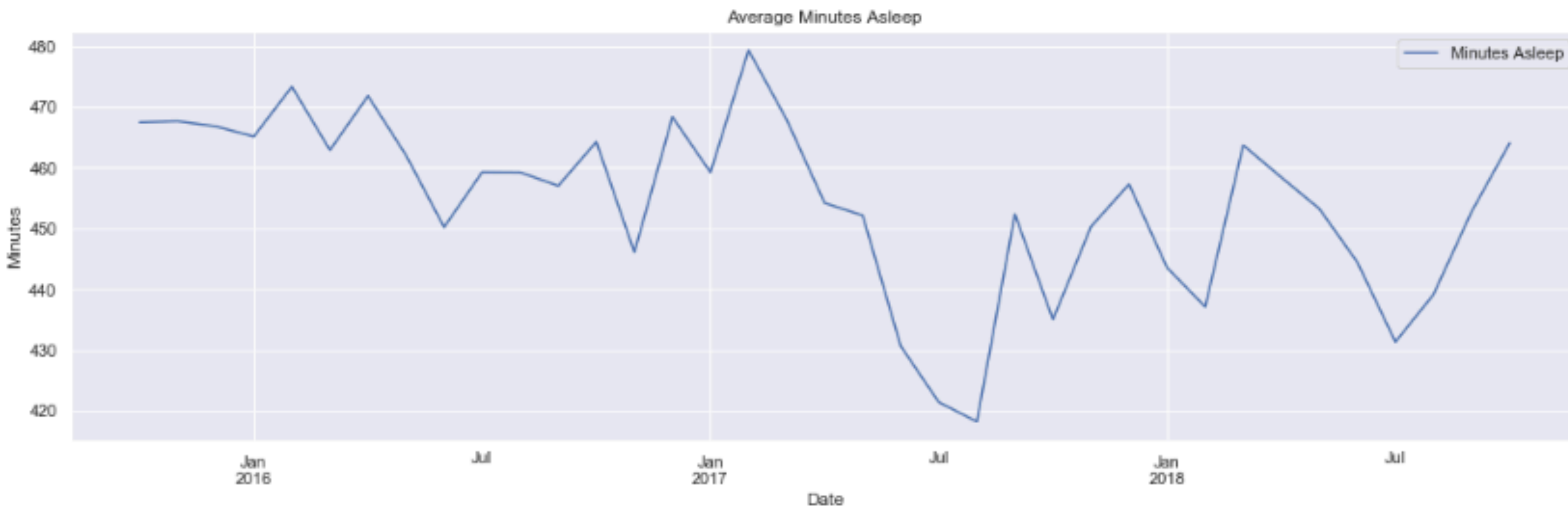
DATA STORYTELLING

Overall average calories burned (aggregated by month)

- **A general upward trend in amount of calories Tracy burns from October 2015 until November 2018**
- **Age is not stopping Tracy from staying active**

DATA STORYTELLING

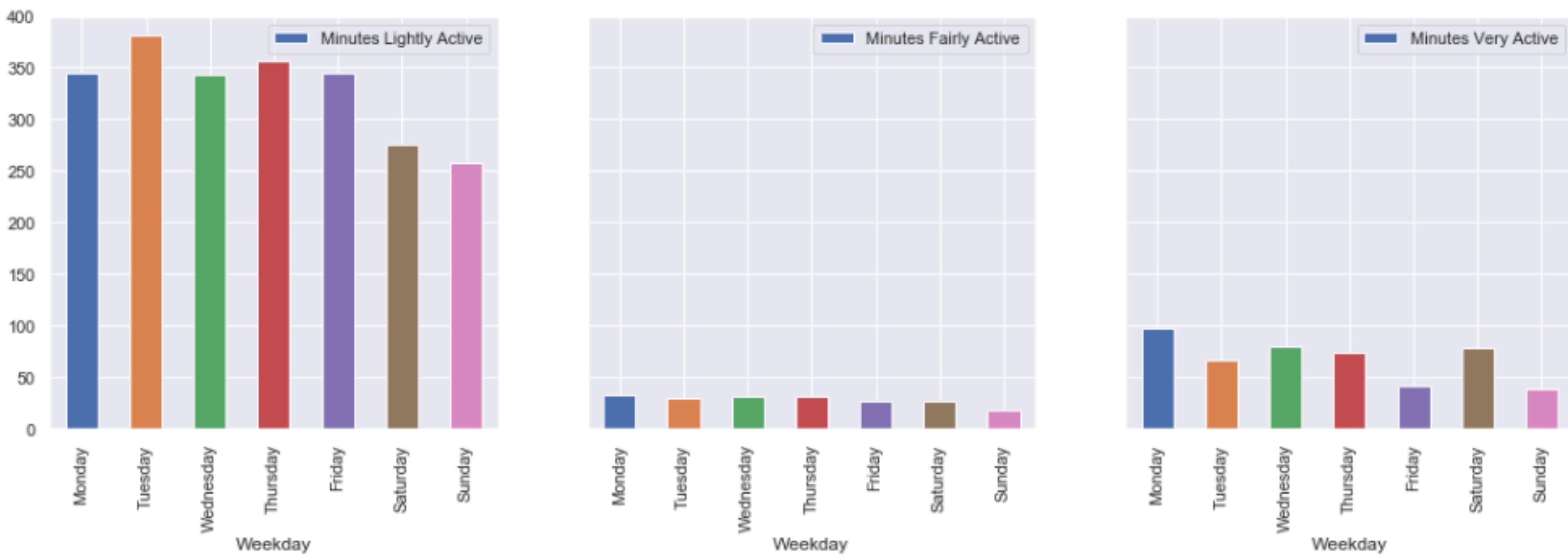
Overall average amount of sleep Tracy gets



- **Unsteady trend**
- **Huge drop from May 2017 to August 2017**
    - A general upward trend afterwards – recovery of sleep time

DATA STORYTELLING

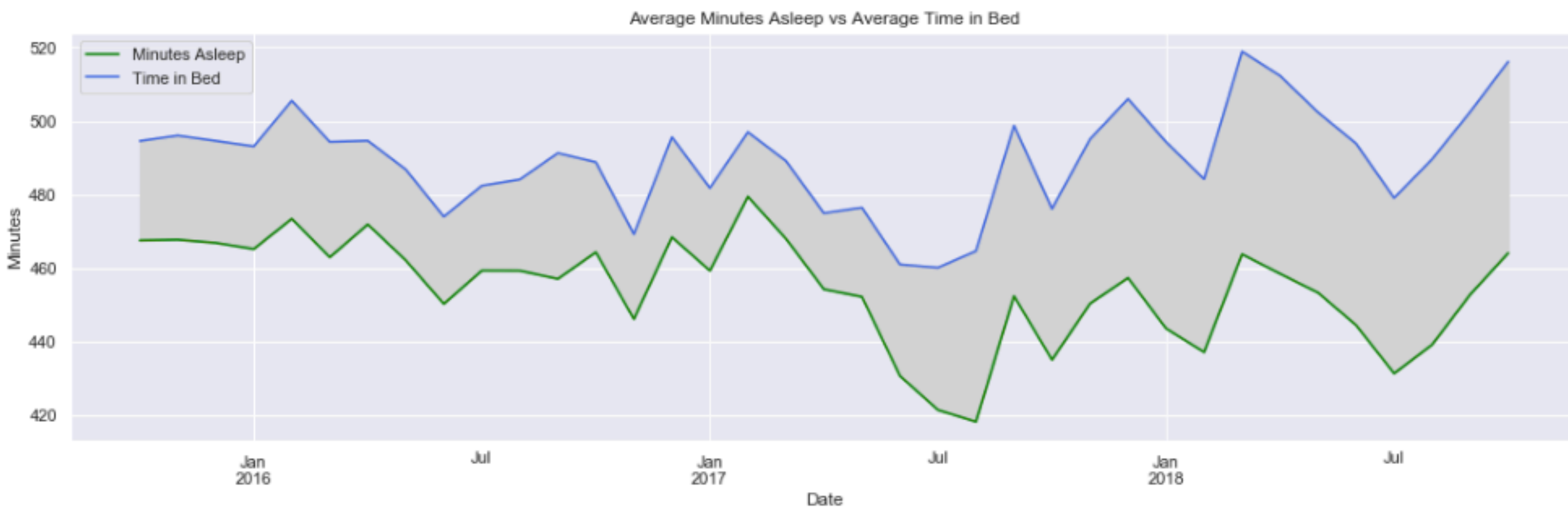
Comparing average minutes per activity level (by week)



- **Most physically active during the weekdays, with exception of Saturday**

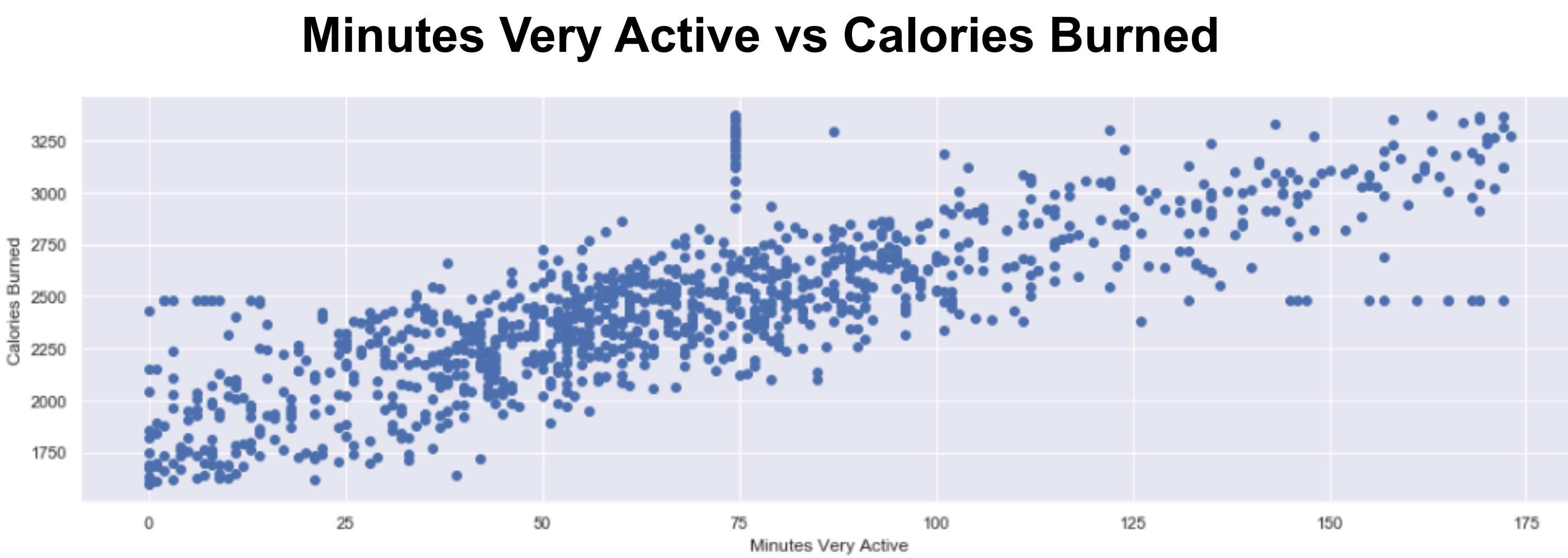
DATA STORYTELLING

Tracy’s restlessness



- **Space between represents Minutes Awake**
- **Since 2017, having more trouble staying asleep**

INFERENTIAL STATS




```python
scipy_r, scipy_p = stats.pearsonr(activities['Minutes Very Active'], \
                                   activities['Calories Burned'])

print("Scipy's correlation coefficient:", scipy_r)
print("Scipy's p-value:", scipy_p)
```

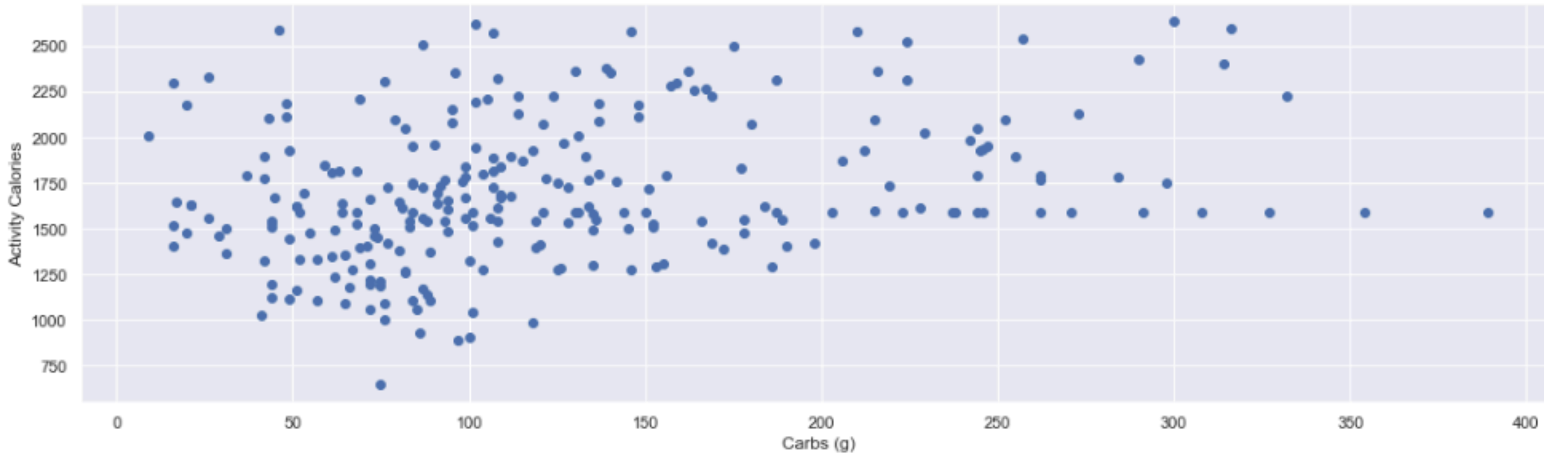
```
Scipy's correlation coefficient: 0.7858467996449217
Scipy's p-value: 7.142083581912041e-233
```

Null hypothesis:
There is no correlation between duration of workout and amount of calories burned

- **Relatively strong, positive correlation**

INFERENTIAL STATS

Does Tracy load up on carbs on her workout intensive days?




```
scipy_r, scipy_p = stats.pearsonr(df1['Activity Calories'], df1['Carbs (g)'])

print("Scipy's correlation coefficient:", scipy_r)
print("Scipy's p-value:", scipy_p)
```

```
Scipy's correlation coefficient: 0.301696397595359
Scipy's p-value: 9.62446731823555e-07
```

Null hypothesis:
There is no correlation between Tracy's activity level and carb intake

- **Weak, but nonetheless positive correlation**
- **p-value suggests rejection of null hypothesis**

MACHINE LEARNING

Through an outside source, heart rate data was also retrieved

	Time	Heart Rate	Date	Time_of_Day
0	00:00:00	57	2017-02-22	Morning
1	00:01:00	55	2017-02-22	Morning
2	00:02:00	50	2017-02-22	Morning
3	00:03:00	51	2017-02-22	Morning
4	00:04:00	51	2017-02-22	Morning
5	00:05:00	51	2017-02-22	Morning
6	00:06:00	52	2017-02-22	Morning
7	00:07:00	51	2017-02-22	Morning
8	00:08:00	51	2017-02-22	Morning
9	00:09:00	51	2017-02-22	Morning
10	00:10:00	52	2017-02-22	Morning

- **‘Date’ and ‘Time_of_Day’ were added via data wrangling**

MACHINE LEARNING

Supervised learning: Predicting calories burned by taking sum of heart rates per day

```
hr_daily_sum.head()
```

Date	Heart Rate	Calories Burned
2017-02-22	96265	2474.0
2017-02-23	100505	2963.0
2017-02-24	93022	2449.0
2017-02-25	94511	2649.0
2017-02-26	93240	2640.0

Chose Linear Regression & Random Forest Regressor models

70% of data used to train model – remaining 30% left to test model's performance

PREDICTING CALORIES BURNED BY TAKING SUM OF HEART RATES PER DAY

Results:

```
LINEAR REGRESSION
-----------------
Mean absolute error: 234.7799212462273
Mean error: 297.1541601692999

Cross validation results: [0.33977675 0.45372836 0.53440749 0.52467865 0.04831244]


RANDOM FOREST REGRESSOR
-----------------------
Mean absolute error: 231.14511290322582
Mean error: 303.53668597908666

Cross validation results: [-0.2258274   0.3819131   0.41860917  0.51159263  0.22444065]
```

Linear Regression model yields an average prediction error of 234-297 calories

Random Forest Regressor model yields an average prediction error of 231-303 calories

MACHINE LEARNING

Include time of day (morning, afternoon, evening) and take average heart rate

Date	(Heart Rate, Afternoon)	(Heart Rate, Evening)	(Heart Rate, Morning)	Calories Burned
2017-02-22	67.763889	63.197222	69.084388	2474.0
2017-02-23	78.405556	58.660167	72.344633	2963.0
2017-02-24	70.273239	61.425714	69.206538	2449.0
2017-02-25	71.941341	60.113889	69.388807	2649.0
2017-02-26	69.534483	60.157233	69.322222	2640.0

PREDICTING CALORIES BURNED BY AVERAGE HEART RATES BY TIME OF DAY

Can we create a model with better predictive performance?

```
LINEAR REGRESSION
-----------------
Mean absolute error: 212.88384694797185
Mean error: 265.31942766617124

Cross validation results: [ 0.1959427   0.46651428  0.66728062  0.5962532  -0.09702459]

RANDOM FOREST REGRESSOR
-----------------------
Mean absolute error: 181.615986245884
Mean error: 224.7414679986225

Cross validation results: [0.05213838 0.56110478 0.73685765 0.6782538  0.4419277 ]
```

Both models produce smaller predictive errors, especially the Random Forest Regressor model

MACHINE LEARNING

Unsupervised Learning

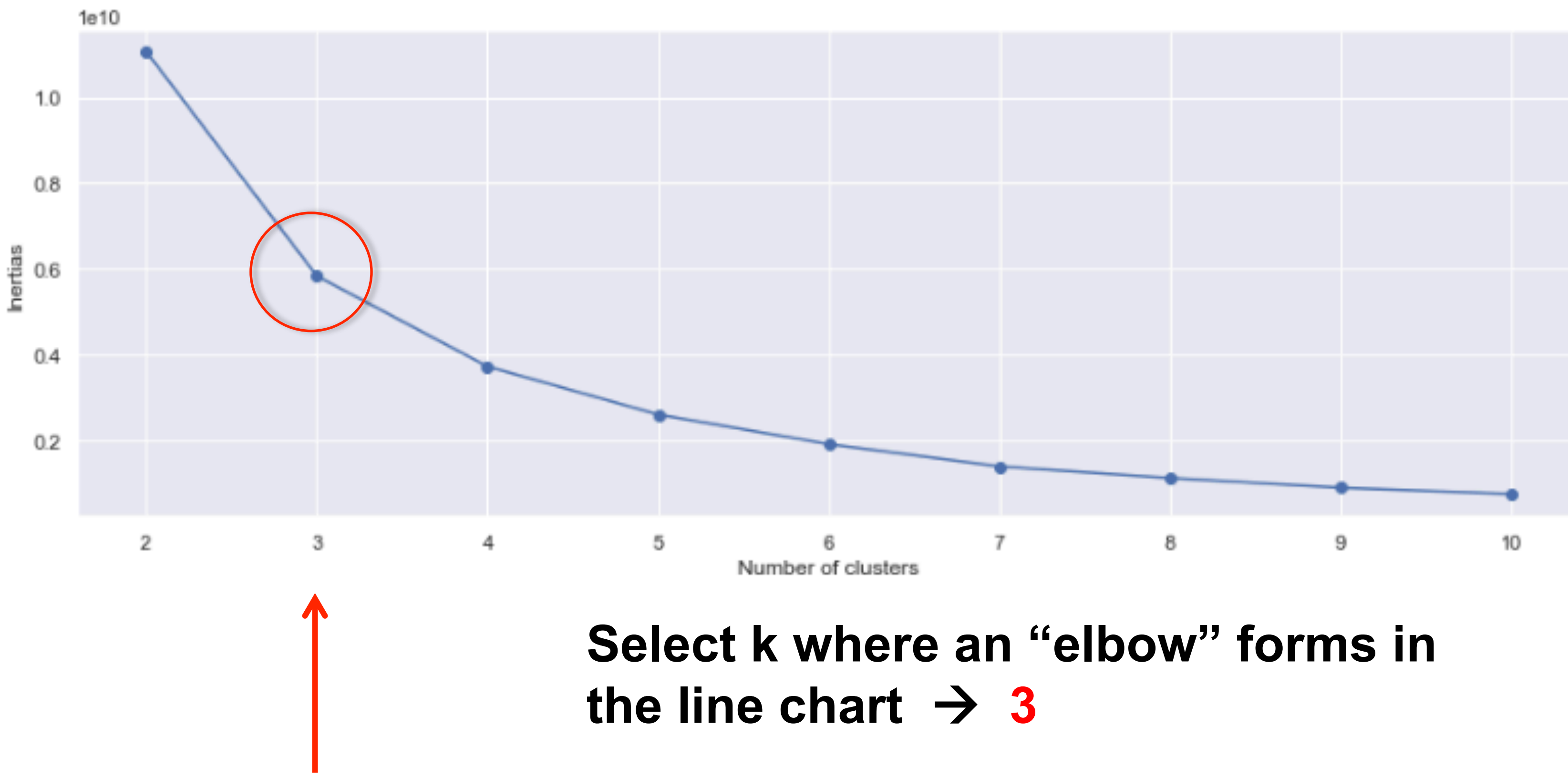
- No labeled data
- Uncover patterns in data

Chosen method: K-Means Clustering

- Group similar data points together an discover underlying patterns

K-MEANS CLUSTERING

Choosing K (number of clusters/groups) to optimize clustering results



Select k where an "elbow" forms in the line chart → 3

UNSUPERVISED LEARNING

How are these clusters formed?

- **By activities? Sleep? Heart rate?**
- **Analyze data to find patterns**

UNSUPERVISED LEARNING

```
CLUSTER 1:
Monday        51
Wednesday     46
Saturday      22
Tuesday       14
Friday         7
Thursday       6
Sunday         2
Name: Weekday, dtype: int64
```

```
CLUSTER 2:
Thursday      63
Tuesday       62
Friday        36
Saturday      34
Wednesday     34
Monday        30
Sunday        13
Name: Weekday, dtype: int64
```

```
CLUSTER 0:
Sunday        72
Friday        35
Saturday      30
Thursday      15
Tuesday       12
Wednesday      7
Monday         5
Name: Weekday, dtype: int64
```

- **Through investigation, 'Weekday' seems to be most deterministic of how clusters are formed**
  - Cluster 0: Sunday's
  - Cluster 1: Monday's and Wednesday's
  - Cluster 2: Tuesday's and Thursday's

…something comes to mind

CLUSTER 1:

Monday	51
Wednesday	46
Saturday	22
Tuesday	14
Friday	7
Thursday	6
Sunday	2

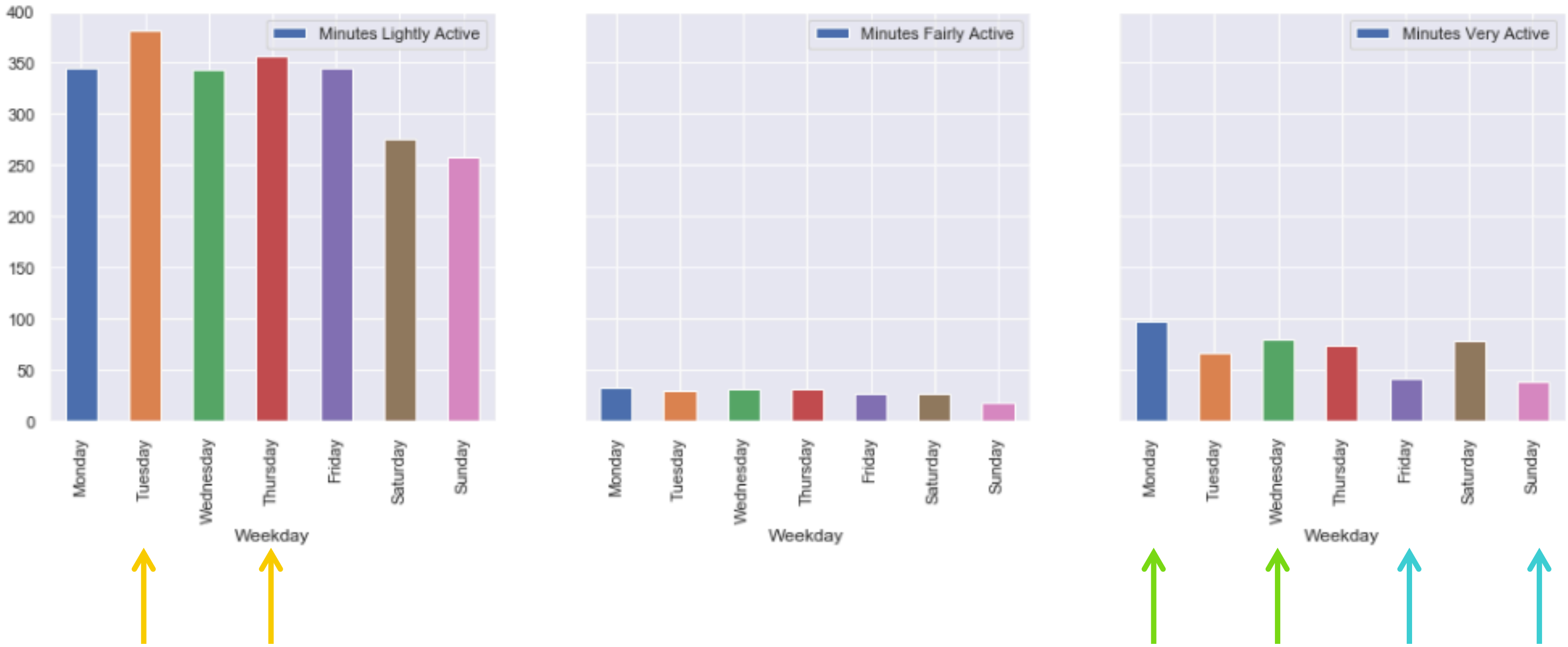
CLUSTER 2:

Thursday	63
Tuesday	62
Friday	36
Saturday	34
Wednesday	34
Monday	30
Sunday	13

CLUSTER 0:

Sunday	72
Friday	35
Saturday	30
Thursday	15
Tuesday	12
Wednesday	7
Monday	5

4



The top two days of week of each cluster seem to have something in common

```
MINUTES SEDENTARY
-----------------
Cluster 0: 536.8065476190476
Cluster 1: 355.4545454545454
Cluster 2: 454.8181818181818

MINUTES LIGHTLY ACTIVE
----------------------
Cluster 0: 291.50119590211057
Cluster 1: 338.4547474303572
Cluster 2: 343.1639898506828

MINUTES FAIRLY ACTIVE
---------------------
Cluster 0: 21.869318181818183
Cluster 1: 40.34820186039698
Cluster 2: 34.39059261916149

MINUTES VERY ACTIVE
-------------------
Cluster 0: 40.26136363636363
Cluster 1: 114.07935984765253
Cluster 2: 81.99275665550773
```

- Cluster 0 has highest mean under **Minutes Sedentary** category
- Cluster 1 has highest mean under **Minutes Very Active** category
- Cluster 2 has a slightly higher mean under **Minutes Lightly Active** category

So it seems the **clusters** are formed by the **days of the week**, which in turn says a lot about **how active** Tracy is on those days

CONCLUSION

- **Recommendations**
  - With evidence of increasing restless nights, Tracy should seek ways to obtain better, uninterrupted sleep
  - There is some evidence of a positive correlation between carb intake and activeness. Tracy should consider fueling her body with carbs to have a more productive workout session.
- **Next steps**
  - Using heart rate data to perform predictions on sleep patterns
    - Can heart rate be used to predict sleep stage?
  - Creating more features based on existing data to detect more patterns in activities and/or sleep
    - Split data by season