Data Journey

ข้อมูลชุดนี้เป็นข้อมูลผู้เสียชีวิตจากอุบัติเหตุบนท้องถนน โดยมีแหล่งข้อมูลจากสำนักงานสถิติแห่งชาติ จำนวนทั้งสิ้น 21 ตัวแปร จำนวนมากกว่า 2 แสนข้อมูล ซึ่งจะประกอบด้วยรายละเอียดของอุบัติเหตุและเสียชีวิตตั้งแต่ปีพ.ศ. 2554 ถึง 2565 (สิ้นสุดวันที่ 3 มิถุนายน) โดยตั้งสมมติฐานหรือคำถามที่ผู้จัดทำสงสัยคือ ระหว่างรถจักรยานยนต์กับรถยนต์ พาหนะใดก่อให้เกิดอุบัติหตุและเสียชีวิตมากกว่ากัน โดยผู้จัดทำคาดเดาว่าน่าจะเป็นรถจักรยานยนต์มากกว่า จึงทำการทดสอบ โดยมีขั้นตอนดังนี้

ในขั้นตอนที่ 1) นำข้อมูลมาดูภาพรวมเพื่อดูรายละเอียด พร้อมทั้งตัวแปรที่สามารถใช้งานได้ ในที่นี้สนใจตัวแปรคือ อายุ เพศ สถานที่ประสบอุบัติเหตุและเสียชีวิต ปีที่เสียชีวิต และยานพาหนะที่ประสบอุบัติเหตุ ซึ่งจากตัวแปรที่กล่าวมา มีค่า null อยู่ค่อนข้างเยอะ จึงทำการ clean data ดังภาพ

```
In [2]:
#import
df_datadead = pd.read_csv("tbl_rtddi_filter_data.csv")

df_datadead

C:\Users\WINDOWS\anaconda3\lib\site-packages\IPython\core\interactiveshell.py:3444: DtypeWarning: Columns (5,6,7,8,12,13) have mixed types.Specify dtype option on import or set low_memory=False.
  exec(code_obj, self.user_global_ns, self.user_ns)
```

Out[2]:

	id	DEAD_YEAR(Budha)	DEAD_YEAR	Age	Sex	Nationalityld	Tumbol	District	Province	RiskHelmet	...	AccSubDist	AccDist	Acc
0	8596637	2554	2011	16.0	1.0	99.0	คลองขนาน	เหนือคลอง	กบ	NaN	...	คลองขนาน	เหนือคลอง	
1	8596636	2554	2011	14.0	1.0	99.0	เกาะศรีบอยา	เหนือคลอง	กบ	NaN	...	คลองขนาน	เหนือคลอง	
2	8609447	2554	2011	NaN	1.0	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	
3	8603763	2554	2011	NaN	1.0	NaN	ปลายพระยา	ปลายพระยา	กบ	NaN	...	ปลายพระยา	ปลายพระยา	
4	8596708	2554	2011	38.0	1.0	99.0	เขาทอง	เมืองกระบี่	กบ	NaN	...	ทับปริก	เมืองกระบี่	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
232164	11417633	2565	2022	16.0	1.0	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	อุบลร
232165	11417640	2565	2022	62.0	1.0	Thai	NaN	สำโรง	อุบลราชธานี	NaN	...	สำโรง	สำโรง	อุบลร
232166	11417635	2565	2022	29.0	1.0	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	อุบลร
232167	11422569	2565	2022	23.0	1.0	Thai	NaN	NaN	NaN	NaN	...	กลางใหญ่	เขื่องใน	อุบลร
232168	11417626	2565	2022	16.0	1.0	Thai	NaN	เขื่องใน	อุบลราชธานี	NaN	...	เขื่องใน	เขื่องใน	อุบลร

ทำการ import data เข้ามา

```
df_select = df_datadead.loc[:,['DEAD_YEAR','Age','Sex','AccProv','Vehicle','MonthDate']]

df_selectyear = df_select.loc [ (df_select['DEAD_YEAR'] >= 2018) & (df_select['DEAD_YEAR'] >= 2018), : ]
df_selectvehicle = df_selectyear.loc [ (df_selectyear['Vehicle'] == 'รถจักรยานยนต์') | (df_selectyear['Vehicle']=='รถยนต์' ) ]
#select_year = df_select.loc[:,(df_select['DEAD_YEAR'])>=2018 & ['DEAD_YEAR']<=2022]
df_selectvehicle.isnull().sum()   #check null
```

```
DEAD_YEAR      0
Age          607
Sex         3356
AccProv        0
Vehicle        0
MonthDate      0
dtype: int64
```

Check null เมื่อพบว่ามีค่า null ในตัวแปรที่จะนำมาวิเคราะห์ จึงทำการ

```
df = df_selectvehicle.dropna()
df.isnull().sum()    #ค่า null เป็น 0

#_____________________ clean ok _____________#
```

```
DEAD_YEAR    0
Age          0
Sex          0
AccProv      0
Vehicle      0
MonthDate    0
dtype: int64
```

ขั้นตอนที่ 2) เมื่อทำการ clean data ได้ตามที่ต้องการแล้ว จึงนำข้อชุดมูลที่เลือกมา แปลงค่าเป็นข้อมูลที่มีลักษณะอ่านง่ายยิ่งขึ้น โดยทำการแปลงเพศจาก ตัวเลข 1,2 เป็นเพศชาย(Male) และ เพศหญิง(Female) ตามลำดับ รวมถึงแปลงเดือนจากตัวเลขเป็นตัวหนังสือตามลำดับเดือนสากล ได้ดังนี้

In [5]:

```
df.loc[ df['Sex']==1.0 , 'Sex' ] = 'Male'
df.loc[ df['Sex']==2.0 , 'Sex' ] = 'Female'

month = ['January','February','March','April','May','June','July','August','September','October','November','December']
df.loc[ df['MonthDate']== 1  , 'MonthDate'] = 'January'
df.loc[df['MonthDate']== 2  , 'MonthDate'] = 'February'
df.loc[df['MonthDate']== 3  , 'MonthDate'] = 'March'
df.loc[df['MonthDate']== 4  , 'MonthDate'] = 'April'
df.loc[df['MonthDate']== 5  , 'MonthDate'] = 'May'
df.loc[df['MonthDate']== 6  , 'MonthDate'] = 'June'
df.loc[df['MonthDate']== 7  , 'MonthDate'] = 'July'
df.loc[df['MonthDate']== 8  , 'MonthDate'] = 'August'
df.loc[df['MonthDate']== 9  , 'MonthDate'] = 'September'
df.loc[df['MonthDate']== 10  , 'MonthDate'] = 'October'
df.loc[df['MonthDate']== 11  , 'MonthDate'] = 'November'
df.loc[df['MonthDate']== 12  , 'MonthDate'] = 'December'

df
```

```
C:\Users\WINDOWS\anaconda3\lib\site-packages\pandas\core\indexing.py:1817: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  self._setitem_single_column(loc, value, pi)
```

Out[5]:

	DEAD_YEAR	Age	Sex	AccProv	Vehicle	MonthDate
148922	2018	9.0	Male	กระบี่	รถยนต์	January
148923	2018	20.0	Male	กระบี่	รถจักรยานยนต์	January
148924	2018	25.0	Female	กระบี่	รถจักรยานยนต์	January
148925	2018	29.0	Male	กระบี่	รถจักรยานยนต์	January
148926	2018	18.0	Male	กระบี่	รถจักรยานยนต์	January

	DEAD_YEAR	Age	Sex	AccProv	Vehicle	MonthDate
148922	2018	9.0	Male	กระบี่	รถยนต์	January
148923	2018	20.0	Male	กระบี่	รถจักรยานยนต์	January
148924	2018	25.0	Female	กระบี่	รถจักรยานยนต์	January
148925	2018	29.0	Male	กระบี่	รถจักรยานยนต์	January
148926	2018	18.0	Male	กระบี่	รถจักรยานยนต์	January
...	...	...	...	...	...	...
232160	2022	21.0	Male	อุบลราชธานี	รถจักรยานยนต์	June
232163	2022	67.0	Male	อุบลราชธานี	รถจักรยานยนต์	June
232165	2022	62.0	Male	อุบลราชธานี	รถจักรยานยนต์	June
232166	2022	29.0	Male	อุบลราชธานี	รถจักรยานยนต์	June
232167	2022	23.0	Male	อุบลราชธานี	รถจักรยานยนต์	June

46583 rows × 6 columns

ขั้นตอนที่ 3) เป็นลำดับของการวิเคราะห์ข้อมูล โดยเริ่มจาก groupby ตัวแปรที่เราต้องการจะวิเคราะห์ ดังนี้

In [6]:
```
#2018s
select_year2018 = df.loc[df["DEAD_YEAR"]==2018]

groupby_year2018 = select_year2018.groupby(['DEAD_YEAR'])
groupby_sex2018 = select_year2018.groupby(['Sex'])
groupby_vehicle2018 = select_year2018.groupby(['Vehicle'])
groupby_month2018 = select_year2018.groupby(['MonthDate'])

count_year2018 = groupby_year2018.size()
count_sex2018 = groupby_sex2018.size()
count_vehicle2018 = groupby_vehicle2018.size()
count_month2018 = groupby_month2018.size()

print(f"{count_sex2018}\n\n{count_vehicle2018}\n\n{count_month2018}")
```

```
Sex
Female    2186
Male      8466
dtype: int64

Vehicle
รถจักรยานยนต์    9580
รถยนต์           1072
dtype: int64

MonthDate
April         988
August        764
December      924
February      998
January      1037
July          780
June          805
```


ซึ่งจะทำในลักษณะนี้จำนวน 5 ปีย้อนหลัง จะพบรายละเอียดจากการ groupby และ count ข้อมูลดังนี้

ในปี 2018

```
Sex
Female    2186
Male      8466
dtype: int64

Vehicle
รถจักรยานยนต์    9580
รถยนต์           1072
dtype: int64

MonthDate
April         988
August        764
December      924
February      998
January      1037
July          780
June          805
March        1042
May           853
November      932
October       770
September     759
dtype: int64
```

ในปี 2019

```
Sex
Female    2298
Male      8679
dtype: int64

Vehicle
รถจักรยานยนต์    9948
รถยนต์           1029
dtype: int64

MonthDate
April         1009
August         837
December       843
February       966
January       1075
July           885
June           940
March         1039
May            875
November       848
October        847
September      813
dtype: int64
```

ในปี 2020

```
Sex
Female    2273
Male      9128
dtype: int64

Vehicle
รถจักรยานยนต์    9194
รถยนต์           2207
dtype: int64

MonthDate
April          579
August         950
December      1171
February      1026
January       1055
July           928
June           822
March         1050
May            837
November      1094
October        986
September      903
dtype: int64
```

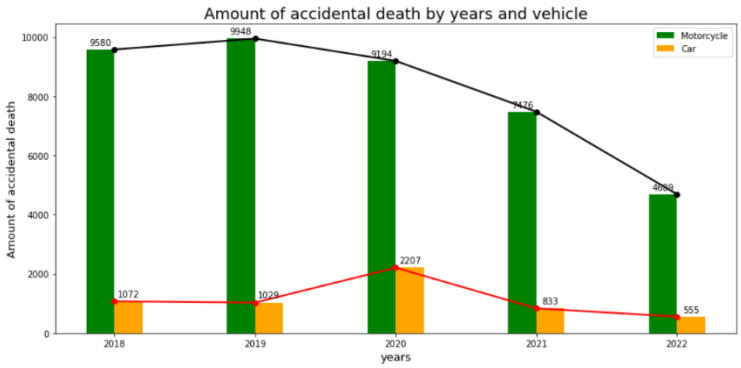
ในปี 2021

```
Sex
Female    1734
Male      6575
dtype: int64

Vehicle
รถจักรยานยนต์    7476
รถยนต์            833
dtype: int64

MonthDate
April         726
August        584
December      822
February      807
January       688
July          588
June          590
March         908
May           651
November      787
October       631
September     527
dtype: int64
```

ในปี 2022

```
Sex
Female    1158
Male      4086
dtype: int64

Vehicle
รถจักรยานยนต์    4689
รถยนต์            555
dtype: int64

MonthDate
April         920
February      832
January       925
June          809
March         902
May           856
dtype: int64
```

ขั้นตอนที่ 4) เป็นการทำ **Visualization** ที่จะนำสิ่งที่เราวิเคราะห์ได้ มา **show** ให้เห็นภาพมากยิ่งขึ้น โดยทางผู้จัดทำได้ทำ นำข้อมูลผู้ประสบอุบัติเหตุและเสียชีวิตรายปีและจำแนกตามพาหนะ(รถยนต์และรถจักรยานยนต์) มานำเสนอ จะได้ลักษณะดังนี้



ในระหว่างการจัดได้พบว่า ช่วงอายุของผู้เสียชีวิตมาในลักษณะข้อมูลดิบ จึงจัดทำ คอลัมน์ดัมมี่ขึ้นมาชื่อว่า Age_group ดังภาพ

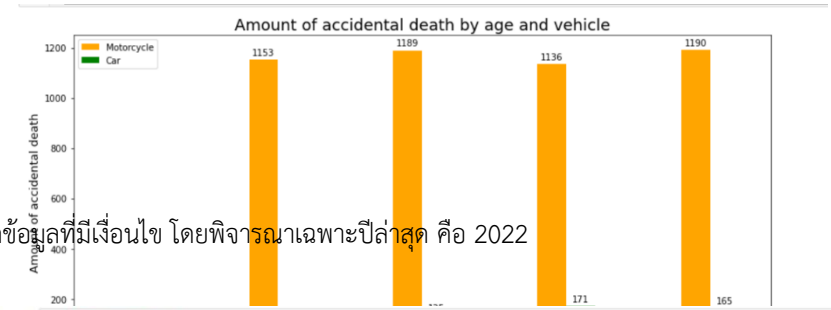
In [16]:
```
# Add a dummy column containing an empty string
df[ 'Age_group' ] = ''
df
#จำแนกอายุให้อยู่ตามกลุ่มต่างๆ

select_year22 = df.loc[df["DEAD_YEAR"]==2022]
select_year22.loc[ (df['Age'] >= 0.0)& (df['Age'] <= 8.0),'Age_group'] = 'ช่วงอายุไม่เกิน 8 ปี'
select_year22.loc[ (df['Age'] >= 9.0)& (df['Age'] <= 24.0),'Age_group'] = 'อายุ 9-24 ปี'
select_year22.loc[ (df['Age'] >= 25.0)& (df['Age'] <= 40.0),'Age_group'] = 'อายุ 25-40 ปี'
select_year22.loc[ (df['Age'] >= 41.0)& (df['Age'] <= 56.0),'Age_group'] = 'อายุ 41-56 ปี'
select_year22.loc[ (df['Age'] >= 57.0) ,'Age_group'] = 'ช่วงอายุมากกว่า 56 ปี'

select_year22

```

5]:

	DEAD_YEAR	Age	Sex	AccProv	Vehicle	MonthDate	Age_group
223545	2022	56.0	Male	กระบี่	รถจักรยานยนต์	January	อายุ 41-56 ปี
223546	2022	66.0	Male	กระบี่	รถจักรยานยนต์	January	ช่วงอายุมากกว่า 56 ปี
223547	2022	21.0	Male	กระบี่	รถยนต์	January	อายุ 9-24 ปี
223549	2022	22.0	Male	กระบี่	รถจักรยานยนต์	January	อายุ 9-24 ปี
223550	2022	58.0	Male	กระบี่	รถจักรยานยนต์	January	ช่วงอายุมากกว่า 56 ปี
...	...	...	...	...	...	...	...
232160	2022	21.0	Male	อุบลราชธานี	รถจักรยานยนต์	June	อายุ 9-24 ปี
232163	2022	67.0	Male	อุบลราชธานี	รถจักรยานยนต์	June	ช่วงอายุมากกว่า 56 ปี
232165	2022	62.0	Male	อุบลราชธานี	รถจักรยานยนต์	June	ช่วงอายุมากกว่า 56 ปี
232166	2022	29.0	Male	อุบลราชธานี	รถจักรยานยนต์	June	อายุ 25-40 ปี
232167	2022	23.0	Male	อุบลราชธานี	รถจักรยานยนต์	June	อายุ 9-24 ปี

5244 rows × 7 columns

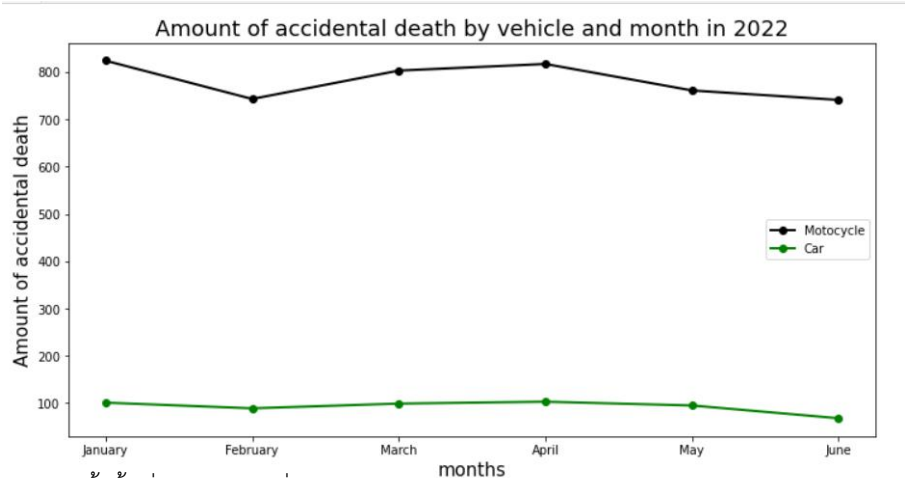
เมื่อนำตัวแปรที่พึ่งสร้างมาใหม่มา ทำการสร้าง bar chart จะได้



และสร้างทำการเลือกข้อมูลที่มีเงื่อนไข โดยพิจารณาเฉพาะปีล่าสุด คือ 2022

```
In [14]:
#วิเคราะห์เฉพาะปี 2022
#โดยวิเคราะห์รายเดือนจำแนกตามยานพาหนะ
select_year22 = df.loc[df["DEAD_YEAR"]==2022]
motor_jan22 = select_year22.loc[(select_year22['Vehicle']=='รถจักรยานยนต์')&(select_year22['MonthDate']=='January')]
car_jan22 = select_year22.loc[(select_year22['Vehicle']=='รถยนต์')&(select_year22['MonthDate']=='January')]
count_jan_motor = motor_jan22.count()
count_jan_car = car_jan22.count()

motor_feb22 = select_year22.loc[(select_year22['Vehicle']=='รถจักรยานยนต์')&(select_year22['MonthDate']=='February')]
car_feb22 = select_year22.loc[(select_year22['Vehicle']=='รถยนต์')&(select_year22['MonthDate']=='February')]
count_feb_motor = motor_feb22.count()
count_feb_car = car_feb22.count()

motor_march22 = select_year22.loc[(select_year22['Vehicle']=='รถจักรยานยนต์')&(select_year22['MonthDate']=='March')]
car_march22 = select_year22.loc[(select_year22['Vehicle']=='รถยนต์')&(select_year22['MonthDate']=='March')]
count_march_motor = motor_march22.count()
count_march_car = car_march22.count()

motor_april22 = select_year22.loc[(select_year22['Vehicle']=='รถจักรยานยนต์')&(select_year22['MonthDate']=='April')]
car_april22 = select_year22.loc[(select_year22['Vehicle']=='รถยนต์')&(select_year22['MonthDate']=='April')]
count_april_motor = motor_april22.count()
count_april_car = car_april22.count()

motor_may22 = select_year22.loc[(select_year22['Vehicle']=='รถจักรยานยนต์')&(select_year22['MonthDate']=='May')]
car_may22 = select_year22.loc[(select_year22['Vehicle']=='รถยนต์')&(select_year22['MonthDate']=='May')]
count_may_motor = motor_may22.count()
count_may_car = car_may22.count()

motor_june22 = select_year22.loc[(select_year22['Vehicle']=='รถจักรยานยนต์')&(select_year22['MonthDate']=='June')]
car_june22 = select_year22.loc[(select_year22['Vehicle']=='รถยนต์')&(select_year22['MonthDate']=='June')]
count_june_motor = motor_june22.count()
count_june_car = car_june22.count()
```

เมื่อนำมา plot เป็น line ดังนี้



ปัญหาที่พบในการทำงานครั้งนี้ เนื่องจากในวันที่ 14 ตุลาคม 2565 เวลาประมาณ 01.55 น. ผู้จัดทำกำลังจะทำ pie chart ที่เป็นส่วนเสริมในแต่ละปี เครื่องโน้ตบุ้คเกิดปัญหาขึ้น เมื่อนำไปซ่อมพบว่า ฮาร์ดดิสเสีย และได้ส่งซ่อม ทำให้ในเช้าวันเดียวกันต้องจัดทำงานขึ้นใหม่หมด **อาจจะมีการแก้ไขงาน เพิ่มในส่วนของกราฟลงไป เพราะในส่วนนี้ทำเพื่อนำเสนอ เป็นการแก้ไขปัญหาเฉพาะหน้าก่อน** รวมถึงพบปัญหาว่าข้อมูลชุดนี้ ในตัวแปรบางมีข้อมูลไม่ครบบ้าง ตอบไม่ตรงกับคำถามบ้าง จึงต้องทำการ drop ข้อมูล

ทิ้งลงไปบางส่วน และเลือกข้อมูลบางส่วนมาทำการวิเคราะห์ รวมถึงปัญหาในการทำ groupby ที่ผู้จัดทำลืมทำการ .size() เพื่อนับจำนวนข้อมูลที่สนใจ