df\_feb = pd.read\_parquet("yellow\_tripdata\_2022-02.parquet")

df\_jun = pd.read\_parquet("yellow\_tripdata\_2022-06.parquet")

df = pd.concat((df\_jun, df\_feb))

df.describe()

df = df[(df['trip\_distance'] > 1) & (df['passenger\_count'] > 0) & (df['fare\_amount']>= 2.5)].dropna().sample(1000000)

#RateCodeId can't be 99 and very few group rides = 6 so removed them as well.

df = df[(df['RatecodeID'] != 99) & (df['RatecodeID'] != 6)]

df = df[df['passenger\_count'] <= 6]

df['pickup\_day'] = df['tpep\_pickup\_datetime'].dt.dayofweek

df['pickup\_weekday'] = df['tpep\_pickup\_datetime'].dt.weekday

df['pickup\_hour'] = df['tpep\_pickup\_datetime'].dt.hour

df['dropoff\_day'] = df['tpep\_dropoff\_datetime'].dt.dayofweek

df['dropoff\_weekday'] = df['tpep\_dropoff\_datetime'].dt.weekday

df['dropoff\_hour'] = df['tpep\_dropoff\_datetime'].dt.hour

df['trip\_month'] = df['tpep\_dropoff\_datetime'].dt.month

df[df['extra'].isin(df['extra'].value\_counts()[df['extra'].value\_counts() > 500].index)]

corr.style.background\_gradient(cmap='coolwarm')

* Used quantile : **0.9995**

| **Categorial** |  | **Remarks** | **Features** | **Temporal** |
| --- | --- | --- | --- | --- |
| Airport Fee |  | Useful variation |  | N |
| Congestion Surcharge |  |  |  | N |
| IMprovement Surcharge |  |  |  | N |
| Vendor ID |  |  |  | N |
| Passenger Count | (0, ) |  |  | N |
| RateCodeID |  |  |  | N |
| Store and fwd | Since, just info about storing data |  |  | N |
| Payment type | [1,1] |  |  | N |
| PU Location ID |  |  |  | Y |
| Du Location ID |  |  |  | Y |
| PIckupday, PIckhour,  Month,  Dropoffday, Dropoffhour | Drop off are highly correlated, make sense | Only pickup hour for temporal dynamics. |  | N |
| Trip\_month = 2.6 |  |  |  | N |
| MTA\_tax |  |  |  | **N** |
| **Continuous** |  |  |  |  |
| trip\_distance | [1,40] | HIghly correlated with total amount | x |  |
| fare\_amount | [2.5,] | HIghly correlated with total amount | Cubic, x^3, x^2, x |  |
| extra(kept continuous even though only few values) |  |  | Power 5, power 4. | Y |
| total\_amount |  |  | **x** | **Y** |
| tolls\_amount |  | Only 1 value for told [60,80]: thus empty bin. | Log, sqrt(x) | Y |
| duration | [,300] | [, 5] | Piecewise[x, x] | Y |
| Tip amount | [ , 200] |  |  |  |

\item Average tip was higher when airport fee was applied.

\item Average tip was higher when there was no congestion surcharge.

\item Average tip roughly varies linearly with trip distance.

\item Average tip was not the same every hour of the day or every day of the week and varied with time.

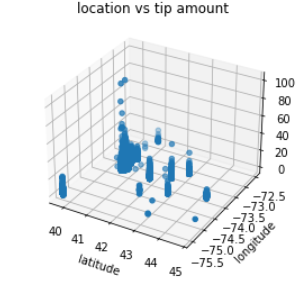
\item Average tip was higher when there was no improvement surcharge.

\item Average tip was higher when MTA tax rate was 0.

\item Average tip varied almost linearly with total charges.

\item Average tip was higher in the month of June as compared to February (No data was extracted from NYC TLC dataset for other months, so the same may not hold for those months)\\

## Old Results without seed fixed

* Linear reg with linear features
  + Extra
    - alpha:10.0, mse:2.4462753018740364, mae:0.7903737168437994
    - alpha:1, mse:2.4462652554901263, mae:0.7903909315445006
    - alpha:0.1, mse:2.4462643095884893, mae:0.790392639062323
    - alpha:0.01, mse:2.446264173758535, mae:0.7903927895913287
    - alpha:0.001, mse:2.4462641837800447, mae:0.7903928454915772
    - alpha:0.0001, mse:2.4462642253997875, mae:0.7903927886302112
    - alpha:1e-05, mse:2.4462642215158725, mae:0.7903928222581408
  + Toll amount
* Independently try each complex features
  + extra^4
    - alpha:10.0, mse:2.4222690223215175, mae:0.7892724786864778
    - alpha:1, mse:2.422254144195369, mae:0.7892922856890541
    - alpha:0.1, mse:2.4222527686337085, mae:0.7892946841648961
    - alpha:0.01, mse:2.422252860359284, mae:0.7892961149698657
    - alpha:0.001, mse:2.4222525557707035, mae:0.7892946574138014
    - alpha:0.0001, mse:2.4222524214815278, mae:0.7892939806731204
    - alpha:1e-05, mse:2.422252567937874, mae:0.7892947337834917
  + extra^5
    - alpha:10.0, mse:2.3860399625446167, mae:0.7904806713635525
    - alpha:1, mse:2.3860616013628926, mae:0.7904971977575683
    - alpha:0.1, mse:2.386063865893186, mae:0.7904989367634706
    - alpha:0.01, mse:2.386064080751092, mae:0.7904991329991562
    - alpha:0.001, mse:2.386064080726063, mae:0.7904991007243953
    - alpha:0.0001, mse:2.3860641830125275, mae:0.7904991821520193
    - alpha:1e-05, mse:2.386064128009692, mae:0.7904991725537589
  + Log(1+toll)
    - alpha:10.0, mse:2.4623314733257993, mae:0.7892959920888548
    - alpha:1, mse:2.462341552246484, mae:0.789311639824574
    - alpha:0.1, mse:2.4623425732571023, mae:0.7893131897234185
    - alpha:0.01, mse:2.4623426737027705, mae:0.7893133505837746
    - alpha:0.001, mse:2.4623426604595453, mae:0.7893133690153082
    - alpha:0.0001, mse:2.4623426966155386, mae:0.7893133416911519
    - alpha:1e-05, mse:2.462342735039337, mae:0.78931335995654
  + sqrt(toll)
    - alpha:10.0, mse:2.39180913439156, mae:0.7888639599829168
    - alpha:1, mse:2.39179996788822, mae:0.7888785975153632
    - alpha:0.1, mse:2.391799115572647, mae:0.7888800696614955
    - alpha:0.01, mse:2.3917990086598127, mae:0.7888801774696023
    - alpha:0.001, mse:2.391799010128501, mae:0.7888802200143916
    - alpha:0.0001, mse:2.391799023593432, mae:0.7888802360908346
    - alpha:1e-05, mse:2.3917990047463342, mae:0.7888802317624213
  + Piecewise duration
    - alpha:10.0, mse:2.396659342560632, mae:0.790891003427179
    - alpha:1, mse:2.3966710718469098, mae:0.7909091409558701
    - alpha:0.1, mse:2.396672260445284, mae:0.7909109966375873
    - alpha:0.01, mse:2.3966724522365723, mae:0.7909111139020932
    - alpha:0.001, mse:2.3966724362847867, mae:0.7909111556192403
    - alpha:0.0001, mse:2.396672410971732, mae:0.7909112086497891
    - alpha:1e-05, mse:2.3966724183450125, mae:0.7909111744980833
* Final with complex best.
  + alpha:10.0, mse:2.5376886014972335, mae:0.793556737581326
  + alpha:1, mse:2.5376758007366944, mae:0.7935746666818139
  + alpha:0.1, mse:2.5376751111332854, mae:0.793576888739156
  + alpha:0.01, mse:2.5376746058704214, mae:0.7935767991638664
  + alpha:0.001, mse:2.5376750846316067, mae:0.7935773190917528
  + alpha:0.0001, mse:2.5376749516751573, mae:0.7935772057331294
  + alpha:1e-05, mse:2.537676618650905, mae:0.7935789677911579
* Hour based b(t), with bins
  + # Week day: [0-1, 2-3, 4, 5-6]
  + Only Hours: [0-3, 4-5, 6-7, 8, 9, 10-12, 13-15, 16, 17, 18, 19, 20, 21-23]
  + # month : [0, 1]
* Extra with hour bins
* Basic + location in x,y form, data\_size: 50000
  + alpha:1000.0, mse:2.125252867592916, mae:0.7911355272155334
  + alpha:100.0, mse:2.1043996194361307, mae:0.7937897766100251
  + alpha:10.0, mse:2.116288404846371, mae:0.7973320936550712
  + alpha:1, mse:2.119401658998576, mae:0.7979526547349872
  + alpha:0.1, mse:2.1198472547426865, mae:0.7983690368587004
  + alpha:0.01, mse:2.119690014738051, mae:0.797993931588696
  + alpha:0.001, mse:2.1197931273250137, mae:0.7980255799779494
  + alpha:0.0001, mse:2.1196853737548333, mae:0.797976847140097
  + alpha:1e-05, mse:2.119985153523083, mae:0.7980878781188516
  + D:S
  + 

# Issues:

* Can’t do one hot encode matrix size explodes, sparse.
  + Even without that PULocationID and DOLocationID sparse is needed

# Results with 60k data

Basic without (pickup\_hour, pickup\_day, trip\_month ) :-

alpha:10.0, mse:2.0610928088736253, mae:0.7826264594058804

alpha:1, mse:2.0708427619118823, mae:0.7867134221583705

alpha:0.1, mse:2.091959247880251, mae:0.7891167666945994

alpha:0.01, mse:2.09695458484354, mae:0.7896417789899546

alpha:0.001, mse:2.097516213508367, mae:0.7896991146038065

alpha:0.0001, mse:2.0975755384233925, mae:0.7897043700299659

alpha:1e-05, mse:2.097601999879929, mae:0.7897219631434448

Basic :-

alpha:10.0, mse:2.0531774923228667, mae:0.7760395757491502

alpha:1, mse:2.062537140132144, mae:0.7794041462081814

alpha:0.1, mse:2.0834575218610376, mae:0.7815980804819812

alpha:0.01, mse:2.088394816226788, mae:0.7821270619025175

alpha:0.001, mse:2.0889605485624267, mae:0.7821860185597965

alpha:0.0001, mse:2.089018955403362, mae:0.7821881159526256

alpha:1e-05, mse:2.0890254418246843, mae:0.7821923774041535

basic + extra^4 instead of extra:-

alpha:10.0, mse:2.0518520516328875, mae:0.7756524689126374

alpha:1, mse:2.061462585727438, mae:0.7790984789226865

alpha:0.1, mse:2.08248929939963, mae:0.7813152344650786

alpha:0.01, mse:2.087462335231373, mae:0.7818373176625939

alpha:0.001, mse:2.0880250858559903, mae:0.7818954182916122

alpha:0.0001, mse:2.088075606668119, mae:0.7818963145670491

alpha:1e-05, mse:2.088084677729684, mae:0.7818987448764192

basic + extra^5 instead of extra:-

alpha:10.0, mse:2.05159550127981, mae:0.7756253929292346

alpha:1, mse:2.061214996111852, mae:0.7790709213925827

alpha:0.1, mse:2.0822447908516954, mae:0.7812854953468616

alpha:0.01, mse:2.0872222427252884, mae:0.7818100623861763

alpha:0.001, mse:2.087778505802967, mae:0.78186342841504

alpha:0.0001, mse:2.0878373159840655, mae:0.7818696271389812

alpha:1e-05, mse:2.0878466600185215, mae:0.7818731230106681

basic + log(1+toll) instead of toll:-

alpha:10.0, mse:2.041603713089311, mae:0.7739671732912872

alpha:1, mse:2.049275898378671, mae:0.7768167516052894

alpha:0.1, mse:2.0680471077290674, mae:0.7788625770819713

alpha:0.01, mse:2.0724538518796223, mae:0.7793695504243627

alpha:0.001, mse:2.072956691223507, mae:0.7794285546707912

alpha:0.0001, mse:2.073022792022855, mae:0.7794394642008939

alpha:1e-05, mse:2.073016116066916, mae:0.7794375528404678

basic + sqrt(toll) instead of toll:-

alpha:10.0, mse:2.042512423865251, mae:0.7743253356532758

alpha:1, mse:2.050390044580373, mae:0.7771901254466297

alpha:0.1, mse:2.0694678022958297, mae:0.7792517808169886

alpha:0.01, mse:2.073953291484096, mae:0.7797551723023668

alpha:0.001, mse:2.074472003728897, mae:0.7798133734812828

alpha:0.0001, mse:2.074518837109622, mae:0.7798190020365119

alpha:1e-05, mse:2.074518986996042, mae:0.7798161141798894

basic + Piecewise duration instead of simple duration:-

alpha:10.0, mse:2.0598401544323743, mae:0.7766946591977346

alpha:1, mse:2.071572136138382, mae:0.7801939502374915

alpha:0.1, mse:2.0940451444450585, mae:0.7826099893618758

alpha:0.01, mse:2.0992695352631907, mae:0.7831625791779959

alpha:0.001, mse:2.099855909561831, mae:0.7832239511179473

alpha:0.0001, mse:2.0999280207028157, mae:0.783230361778923

alpha:1e-05, mse:2.099936318115422, mae:0.7832294453209752

basic + extra^5 instead of extra and log(1+toll) instead of toll :-

alpha:10.0, mse:2.039990672014591, mae:0.7735336896945108

alpha:1, mse:2.047889213914557, mae:0.7764646413040606

alpha:0.1, mse:2.066801417292401, mae:0.7785133576034589

alpha:0.01, mse:2.0712536057364983, mae:0.7790260323127924

alpha:0.001, mse:2.0717484019415813, mae:0.7790798520749822

alpha:0.0001, mse:2.07180163346688, mae:0.779081936923077

alpha:1e-05, mse:2.0718097616290496, mae:0.779085782488744

basic + temporal params for (duration, hour bin) pairs :-

alpha:10.0, mse:2.0343010017082883, mae:0.7694008190572315

alpha:1, mse:2.044041812829785, mae:0.7734265582622327

alpha:0.1, mse:2.0658724578878918, mae:0.7757740030921931

alpha:0.01, mse:2.071048501547048, mae:0.7762731228426163

alpha:0.001, mse:2.0716338237207643, mae:0.7763340407225403

alpha:0.0001, mse:2.07169338526792, mae:0.7763393555130665

alpha:1e-05, mse:2.0716993162515718, mae:0.776339205401766

basic + extra^5 and log(1+toll) and temporal params for (duration, hour bin) pairs :-

alpha:10.0, mse:2.0222036273707236, mae:0.767696322753696

alpha:1, mse:2.029862624409823, mae:0.770930981628542

alpha:0.1, mse:2.0491713968867256, mae:0.7731583605658408

alpha:0.01, mse:2.0537342035211776, mae:0.773680287283651

alpha:0.001, mse:2.0542535892363856, mae:0.7737399733488043

alpha:0.0001, mse:2.054304810636632, mae:0.7737459044848393

alpha:1e-05, mse:2.0543074673325323, mae:0.773745865673762

basic + temporal params for (duration, weekday bin) pairs :-

alpha:10.0, mse:2.0541733935160886, mae:0.7771785671064143

alpha:1, mse:2.064149434832908, mae:0.7807402624344626

alpha:0.1, mse:2.085703136161681, mae:0.7829648153212303

alpha:0.01, mse:2.0907544009946863, mae:0.7835120369257051

alpha:0.001, mse:2.0913331025119164, mae:0.7835771610460217

alpha:0.0001, mse:2.0913877462130155, mae:0.7835789520085976

alpha:1e-05, mse:2.0913964767101305, mae:0.7835809312885527

# Results with 2m data

Basic without (pickup\_hour, pickup\_day, trip\_month) :-

alpha:10.0, mse:2.283352926009486, mae:0.7731179215118662

alpha:1, mse:2.2845770377251804, mae:0.7734705827863538

alpha:0.1, mse:2.2859794123252533, mae:0.7736147442940099

alpha:0.01, mse:2.2862002242133626, mae:0.773632680612363

alpha:0.001, mse:2.286228206575827, mae:0.7736354730241403

alpha:0.0001, mse:2.2862139872365392, mae:0.7736373778201115

alpha:1e-05, mse:2.2862197392171595, mae:0.7736400690805

Basic :-

alpha:10.0, mse:2.2683977446493695, mae:0.7671426920789672

alpha:1, mse:2.269594384143705, mae:0.7674297892750976

alpha:0.1, mse:2.270891884943395, mae:0.7675542296482778

alpha:0.01, mse:2.271118729025729, mae:0.7675753947589287

alpha:0.001, mse:2.2711447362632002, mae:0.7675756942331394

alpha:0.0001, mse:2.2711492131926905, mae:0.7675747915782252

alpha:1e-05, mse:2.2711942464474073, mae:0.7675782731553761

Basic + extra^5 + log(1+toll) :-

alpha:10.0, mse:2.274424402514327, mae:0.767190433534476

alpha:1, mse:2.275602295572666, mae:0.7674707877033324

alpha:0.1, mse:2.2769094710212285, mae:0.767603700943584

alpha:0.01, mse:2.277111315230372, mae:0.7676194793931744

alpha:0.001, mse:2.27713299235807, mae:0.7676202476045425

alpha:0.0001, mse:2.2771410110556247, mae:0.767621470184677

alpha:1e-05, mse:2.2771359059922123, mae:0.7676213892990699

Basic + temporal params for (duration, hour bin) pairs :-

alpha:10.0, mse:2.236715059586641, mae:0.7615526911011808

alpha:1, mse:2.2379053654517373, mae:0.761785830257712

alpha:0.1, mse:2.2392891090757794, mae:0.7619126393277148

alpha:0.01, mse:2.239464482610379, mae:0.7619272044281771

alpha:0.001, mse:2.239498824163028, mae:0.7619296874395254

alpha:0.0001, mse:2.239494407470417, mae:0.7619301291628257

alpha:1e-05, mse:2.23949286337022, mae:0.7619310875222478

Basic + extra^5 + log(1+toll) + temporal params for (duration, hour bin) pairs :-

alpha:10.0, mse:2.244066310490338, mae:0.7619662688258109

alpha:1, mse:2.245218916802152, mae:0.7621915757819723

alpha:0.1, mse:2.2466390073735507, mae:0.7623261583819644

alpha:0.01, mse:2.246874131981891, mae:0.7623399286467619

alpha:0.001, mse:2.2468696081390642, mae:0.7623384747732227

alpha:0.0001, mse:2.246874236626824, mae:0.7623409838547114

alpha:1e-05, mse:2.246870965961944, mae:0.7623400548623607

EDA Plots :-

