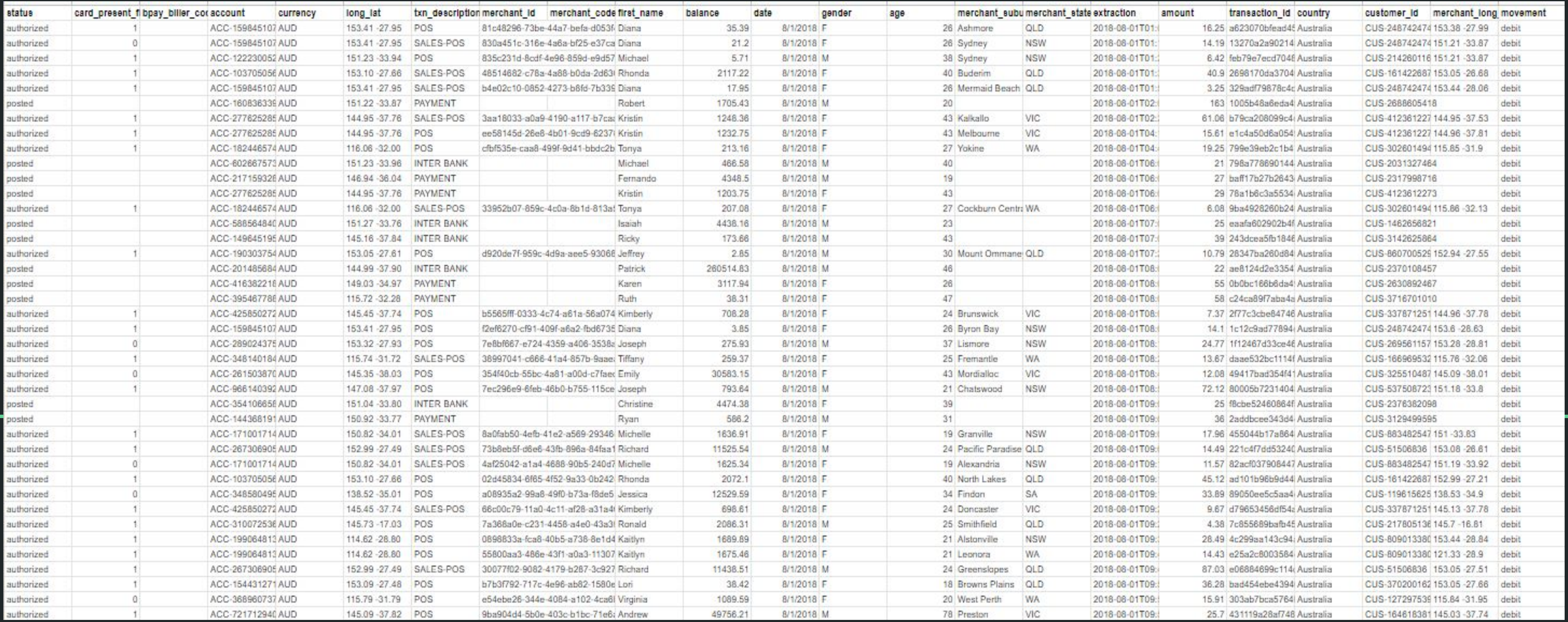
Statistical, Exploratory & Predictive Modelling Findings

Vivek Majithia



status	card_present_f	bpay_biller_co	account	currency	long_lat	txn_description	merchant_id	merchant_code	first_name	balance	date	gender	age	merchant_subu	merchant_state	extraction	amount	transaction_id	country	customer_id	merchant_long	movement
authorized	1		ACC-159845107	AUD	153.41 -27.95	POS	81c48296-73be-44a7-befa-d053f		Diana	35.39	8/1/2018	F	26	Ashmore	QLD	2018-08-01T01:	16.25	a623070bfead45	Australia	CUS-248742474	153.38 -27.99	debit
authorized	0		ACC-159845107	AUD	153.41 -27.95	SALES-POS	830a451c-316e-4a6a-bf25-e37ca		Diana	21.2	8/1/2018	F	26	Sydney	NSW	2018-08-01T01:	14.19	13270a2a90214	Australia	CUS-248742474	151.21 -33.87	debit
authorized	1		ACC-122230052	AUD	151.23 -33.94	POS	835c231d-8cdf-4e96-859d-e9d57		Michael	5.71	8/1/2018	M	38	Sydney	NSW	2018-08-01T01:	6.42	feb79e7ecd7046	Australia	CUS-214260116	151.21 -33.87	debit
authorized	1		ACC-103705056	AUD	153.10 -27.66	SALES-POS	48514682-c78a-4a88-b0da-2d63		Rhonda	2117.22	8/1/2018	F	40	Buderim	QLD	2018-08-01T01:	40.9	2698170da3704	Australia	CUS-161422687	153.05 -26.68	debit
authorized	1		ACC-159845107	AUD	153.41 -27.95	SALES-POS	b4e02c10-0852-4273-b8fd-7b339		Diana	17.95	8/1/2018	F	26	Mermaid Beach	QLD	2018-08-01T01:	3.25	329adf79878c4c	Australia	CUS-248742474	153.44 -28.06	debit
posted			ACC-160836339	AUD	151.22 -33.87	PAYMENT			Robert	1705.43	8/1/2018	M	20			2018-08-01T02:	163	1005b48a6eda4	Australia	CUS-2688605418		debit
authorized	1		ACC-277625285	AUD	144.95 -37.76	SALES-POS	3aa18033-a0a9-4190-a117-b7ca		Kristin	1248.36	8/1/2018	F	43	Kalkallo	VIC	2018-08-01T02:	61.06	b79ca208099c4	Australia	CUS-412361227	144.95 -37.53	debit
authorized	1		ACC-277625285	AUD	144.95 -37.76	POS	ee58145d-26e8-4b01-9cd9-6237		Kristin	1232.75	8/1/2018	F	43	Melbourne	VIC	2018-08-01T04:	15.61	e1c4a50d6a054	Australia	CUS-412361227	144.96 -37.81	debit
authorized	1		ACC-182446574	AUD	116.06 -32.00	POS	cfbf535e-caa8-499f-9d41-bbdc2b		Tonya	213.16	8/1/2018	F	27	Yokine	WA	2018-08-01T04:	19.25	799e39eb2c1b4	Australia	CUS-302601494	115.85 -31.9	debit
posted			ACC-602667573	AUD	151.23 -33.96	INTER BANK			Michael	466.58	8/1/2018	M	40			2018-08-01T06:	21	798a778690144	Australia	CUS-2031327464		debit
posted			ACC-217159328	AUD	146.94 -36.04	PAYMENT			Fernando	4348.5	8/1/2018	M	19			2018-08-01T06:	27	baff17b27b2643	Australia	CUS-2317998716		debit
posted			ACC-277625285	AUD	144.95 -37.76	PAYMENT			Kristin	1203.75	8/1/2018	F	43			2018-08-01T06:	29	78a1b6c3a5534	Australia	CUS-4123612273		debit
authorized	1		ACC-182446574	AUD	116.06 -32.00	SALES-POS	33952b07-859c-4c0a-8b1d-813a		Tonya	207.08	8/1/2018	F	27	Cockburn Centr	WA	2018-08-01T06:	6.08	9ba4928260b24	Australia	CUS-302601494	115.86 -32.13	debit
posted			ACC-588564840	AUD	151.27 -33.76	INTER BANK			Isaiah	4438.16	8/1/2018	M	23			2018-08-01T07:	25	eaafa602902b4f	Australia	CUS-1462656821		debit
posted			ACC-149645195	AUD	145.16 -37.84	INTER BANK			Ricky	173.66	8/1/2018	M	43			2018-08-01T07:	39	243dcea5fb1846	Australia	CUS-3142625864		debit
authorized	1		ACC-190303754	AUD	153.05 -27.61	POS	d920de7f-959c-4d9a-aee5-93068		Jeffrey	2.85	8/1/2018	M	30	Mount Ommane	QLD	2018-08-01T07:	10.79	28347ba260d84	Australia	CUS-860700529	152.94 -27.55	debit
posted			ACC-201485684	AUD	144.99 -37.90	INTER BANK			Patrick	260514.83	8/1/2018	M	46			2018-08-01T08:	22	ae8124d2e3354	Australia	CUS-2370108457		debit
posted			ACC-416382218	AUD	149.03 -34.97	PAYMENT			Karen	3117.94	8/1/2018	F	26			2018-08-01T08:	55	0b0bc166b6da4	Australia	CUS-2630892467		debit
posted			ACC-395467788	AUD	115.72 -32.28	PAYMENT			Ruth	38.31	8/1/2018	F	47			2018-08-01T08:	58	c24ca89f7aba4a	Australia	CUS-3716701010		debit
authorized	1		ACC-425850272	AUD	145.45 -37.74	POS	b5565fff-0333-4c74-a61a-56a074		Kimberly	708.28	8/1/2018	F	24	Brunswick	VIC	2018-08-01T08:	7.37	2f77c3cbe84746	Australia	CUS-337871251	144.96 -37.78	debit
authorized	1		ACC-159845107	AUD	153.41 -27.95	POS	f2ef6270-cf91-409f-a6a2-fbd6735		Diana	3.85	8/1/2018	F	26	Byron Bay	NSW	2018-08-01T08:	14.1	1c12c9ad77894	Australia	CUS-248742474	153.6 -28.63	debit
authorized	0		ACC-289024375	AUD	153.32 -27.93	POS	7e8bf667-e724-4359-a406-3538a		Joseph	275.93	8/1/2018	M	37	Lismore	NSW	2018-08-01T08:	24.77	1f12467d33ce46	Australia	CUS-269561157	153.28 -28.81	debit
authorized	1		ACC-348140184	AUD	115.74 -31.72	SALES-POS	38997041-c666-41a4-857b-9aae		Tiffany	259.37	8/1/2018	F	25	Fremantle	WA	2018-08-01T08:	13.67	daae532bc1114f	Australia	CUS-166969532	115.76 -32.06	debit
authorized	0		ACC-261503870	AUD	145.35 -38.03	POS	354f40cb-55bc-4a81-a00d-c7fae		Emily	30583.15	8/1/2018	F	43	Mordialloc	VIC	2018-08-01T08:	12.08	49417bad354f41	Australia	CUS-325510487	145.09 -38.01	debit
authorized	1		ACC-966140392	AUD	147.08 -37.97	POS	7ec296e9-6feb-46b0-b755-115ce		Joseph	793.64	8/1/2018	M	21	Chatswood	NSW	2018-08-01T08:	72.12	80005b7231404	Australia	CUS-537508723	151.18 -33.8	debit
posted			ACC-354106658	AUD	151.04 -33.80	INTER BANK			Christine	4474.38	8/1/2018	F	39			2018-08-01T09:	25	f8cbe52460864f	Australia	CUS-2376382098		debit
posted			ACC-144368191	AUD	150.92 -33.77	PAYMENT			Ryan	586.2	8/1/2018	M	31			2018-08-01T09:	36	2addbcee343d4	Australia	CUS-3129499595		debit
authorized	1		ACC-171001714	AUD	150.82 -34.01	SALES-POS	8a0fab50-4efb-41e2-a569-29346		Michelle	1636.91	8/1/2018	F	19	Granville	NSW	2018-08-01T09:	17.96	455044b17a864	Australia	CUS-883482547	151 -33.83	debit
authorized	1		ACC-267306905	AUD	152.99 -27.49	SALES-POS	73b8eb5f-d6e6-43fb-896a-84faa1		Richard	11525.54	8/1/2018	M	24	Pacific Paradise	QLD	2018-08-01T09:	14.49	221c4f7dd53240	Australia	CUS-51506836	153.08 -26.61	debit
authorized	0		ACC-171001714	AUD	150.82 -34.01	SALES-POS	4af25042-a1a4-4688-90b5-240d7		Michelle	1625.34	8/1/2018	F	19	Alexandria	NSW	2018-08-01T09:	11.57	82acf037908447	Australia	CUS-883482547	151.19 -33.92	debit
authorized	1		ACC-103705056	AUD	153.10 -27.66	POS	02d45834-6f65-4f52-9a33-0b242		Rhonda	2072.1	8/1/2018	F	40	North Lakes	QLD	2018-08-01T09:	45.12	ad101b96b9d44	Australia	CUS-161422687	152.99 -27.21	debit
authorized	0		ACC-348580495	AUD	138.52 -35.01	POS	a08935a2-99a8-49f0-b73a-f8de5		Jessica	12529.59	8/1/2018	F	34	Findon	SA	2018-08-01T09:	33.89	89050ee5c5aa4	Australia	CUS-119615625	138.53 -34.9	debit
authorized	1		ACC-425850272	AUD	145.45 -37.74	SALES-POS	66c00c79-11a0-4c11-af28-a31a4		Kimberly	698.61	8/1/2018	F	24	Doncaster	VIC	2018-08-01T09:	9.67	d79653456df54a	Australia	CUS-337871251	145.13 -37.78	debit
authorized	1		ACC-310072536	AUD	145.73 -17.03	POS	7a368a0e-c231-4458-a4e0-43a3		Ronald	2086.31	8/1/2018	M	25	Smithfield	QLD	2018-08-01T09:	4.38	7c855689bafb45	Australia	CUS-217805136	145.7 -16.81	debit
authorized	1		ACC-199064813	AUD	114.62 -28.80	POS	0898833a-fca8-40b5-a738-8e1d4		Kaitlyn	1689.89	8/1/2018	F	21	Alstonville	NSW	2018-08-01T09:	28.49	4c299aa143c94	Australia	CUS-809013380	153.44 -28.84	debit
authorized	1		ACC-199064813	AUD	114.62 -28.80	POS	55800aa3-486e-43f1-a0a3-11307		Kaitlyn	1675.46	8/1/2018	F	21	Leonora	WA	2018-08-01T09:	14.43	e25a2c8003584	Australia	CUS-809013380	121.33 -28.9	debit
authorized	1		ACC-267306905	AUD	152.99 -27.49	SALES-POS	30077f02-9082-4179-b287-3c927		Richard	11438.51	8/1/2018	M	24	Greenslopes	QLD	2018-08-01T09:	87.03	e06884699c114	Australia	CUS-51506836	153.05 -27.51	debit
authorized	1		ACC-154431271	AUD	153.09 -27.48	POS	b7b3f792-717c-4e96-ab82-1580e		Lori	38.42	8/1/2018	F	18	Browns Plains	QLD	2018-08-01T09:	36.28	bad454ebe4394	Australia	CUS-370200162	153.05 -27.66	debit
authorized	0		ACC-368960737	AUD	115.79 -31.79	POS	e54ebe26-344e-4084-a102-4ca6		Virginia	1089.59	8/1/2018	F	20	West Perth	WA	2018-08-01T09:	15.91	303ab7bca5764	Australia	CUS-127297539	115.84 -31.95	debit
authorized	1		ACC-721712940	AUD	145.09 -37.82	POS	9ba904d4-5b0e-403c-b1bc-71e6		Andrew	49756.21	8/1/2018	M	78	Preston	VIC	2018-08-01T09:	25.7	431119a28af748	Australia	CUS-164618381	145.03 -37.74	debit

Exploratory Analysis

Exploring Correlations:

- **Shapiro-Wilk test** for normality conducted
    - **Variables** - ‘Amount’, ‘Age’, ‘Balance’
    - *Result - All three samples did not look Gaussian*
- **Pearson correlation** assumptions tested
    - *Result - They had a significant number of outliers*
    - *Result - The relationship between the variables was not approximately linear*
    - *Thus, Pearson test not appropriate*
- Non-parametric **Spearman correlation test** conducted
    - *Result - Weak correlations between each of the variables above.*
- **Chi-Square Test of Association** conducted
    - **Variables** - Gender, Movement, Txn Description
    - *Result - p-value > 5%, fail to reject H0 - No correlation exists between Gender & Movement*
    - *Result - p-value < 5%, reject H0 - Correlation exists between Gender & Txn Description, Cramer V (0.07) proves association is weak.*
    - *Result - p-value < 5%, reject H0 - Correlation exists between Txn Description & Movement, Cramer V (1.00) proves association is strong.*

General Questions (Part 1)

How many transactions do customers make each month on average?

- Average of 4014 transactions per month

What regions had the most transactions by counts?

- Melbourne, Sydney - Top Suburbs
- They also contained more Males than Female transaction makers

What kind of movements were most common?

- Debit movements were more apparent than credits
- Ages 20-30 saw higher debit movements, moderate from ages 30-45 and almost little to none after the age of 45

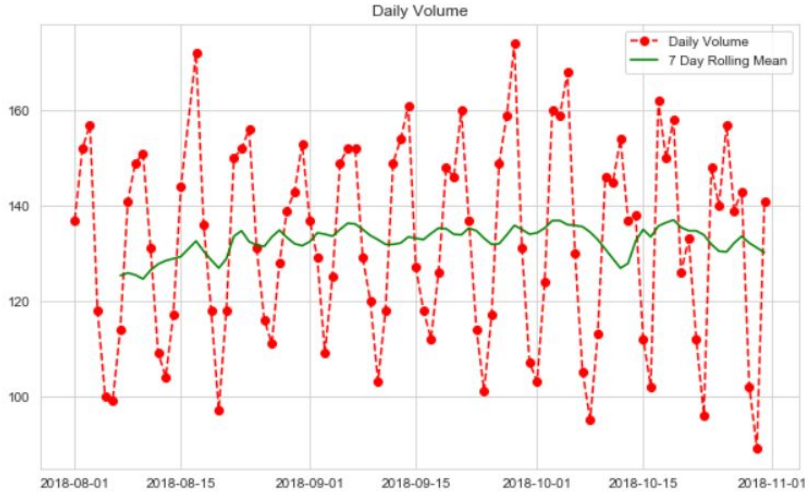
Most common txn_description by gender group?

- POS, SALES-POS, PAYMENT are the most common for both gender groups.

Which month had the highest number of transactions?

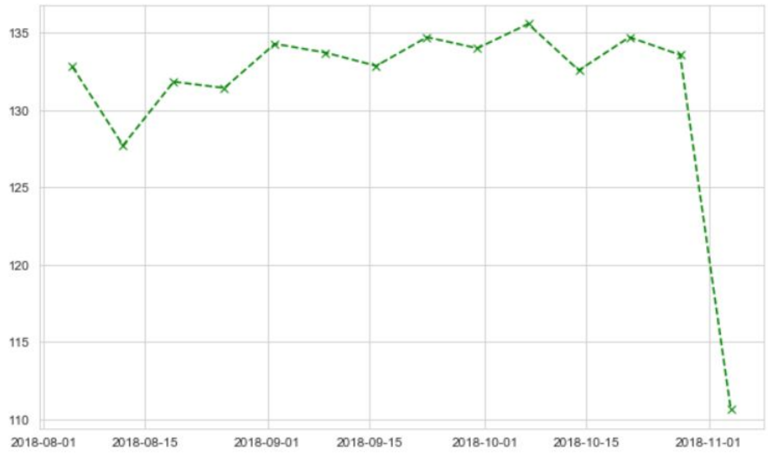
- Month 10 - 4087 total transactions

Transaction Volume per Day



- Maximum of 174 transactions seen in a day which occured on 2018/09/28
- Minimum of 89 transactions seen in a day which occured on 2018/10/30
- Overall average of 132 transactions per day
- Standard deviation of 20 transactions per day

Transaction Volume per Week

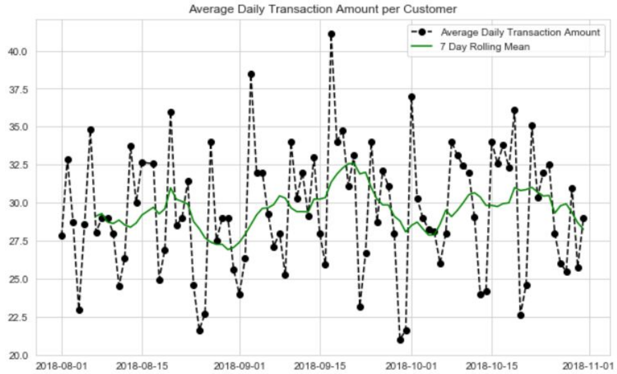


- Progressive highs and lows seen throughout the weeks
- Sudden drop noticed in the last two weeks. More exploration needed

Daily Average Amount per Customer

Amount variable

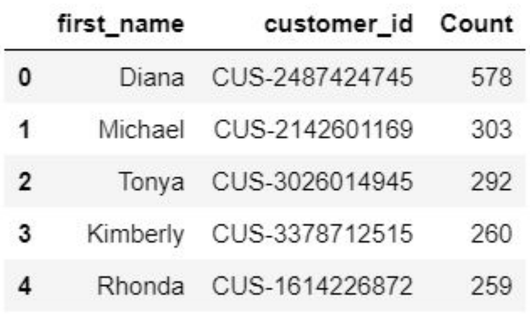
- Right skewed with skewness of 5.35
- Significant outliers present (1357 extreme outliers)
- Median is a better choice of central tendency than the mean in this case
- Median is approximately 29 AUD per Customer



- Maximum daily average amount was 41.09 AUD per customer
- Minimum daily average amount was 21.00 AUD per customer
- The overall daily average was 29.47AUD per customer
- Standard deviation of 4.06 AUD

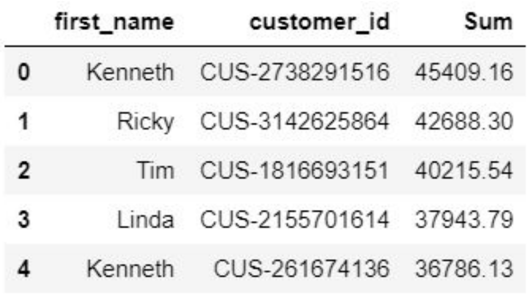
General Questions (Part 2)

Top 5 clients who made the most transactions over 3 months?



	first_name	customer_id	Count
0	Diana	CUS-2487424745	578
1	Michael	CUS-2142601169	303
2	Tonya	CUS-3026014945	292
3	Kimberly	CUS-3378712515	260
4	Rhonda	CUS-1614226872	259

Who was associated with the top 5 largest transactions over 3 months?



	first_name	customer_id	Sum
0	Kenneth	CUS-2738291516	45409.16
1	Ricky	CUS-3142625864	42688.30
2	Tim	CUS-1816693151	40215.54
3	Linda	CUS-2155701614	37943.79
4	Kenneth	CUS-261674136	36786.13

General Questions (Part 3)

Is there a significant difference in average age between the customers that have a debit transaction and those that have a credit transaction?

- Age feature not normally distributed, no important outliers, levene test p-value < 5% thus group variances not equal. Parametric independent sample t-test not possible.
- **Mann-Whitney** test conducted
  - ***Variables*** *- Movement, Age*
  - *Result - p-value < 5%, thus, significant difference in the median of the two groups*

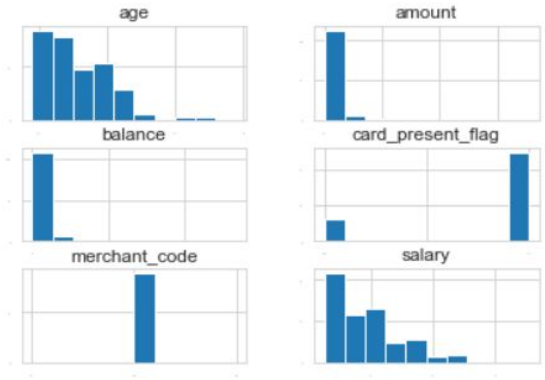
Is there a significant difference between the merchant state groups with respect to the transaction amount?

- Amount feature not normally distributed, contains significant number of outliers, levene test p-value < 5% thus group variances not equal. Parametric one-way ANOVA not possible
- **Kruskal-Wallis** test conducted
  - ***Variables*** *- Amount, Merchant State*
  - *Result - p-value <5%, there is a significant difference between the medians (i.e. average transaction amount) of the 9 groups (8 categories + 1 'Missing' category imputed).*

Predictive Modelling (Regression)

A new feature **Salary** was derived and needed to be predicted. This was the focus of this task

Below shows an example of the distributions for the 6 numerical features



- *Each subplot corresponds to a numerical feature*
- *Some features have a skewed distribution*
- *Standardising or normalising the features may be of value as well as some power transformations to make the distributions appear more Gaussian*
- *Spearman correlation index of 0.508 is seen between the variables **Amount** and **Salary***

Machine Learning Algorithms

Evaluation Metric - **R2 Score**

Method of Data Split - **K-Fold Cross Validation (10 Folds)**

Data Preparation Methods Tested - **Null Values dropped, Standardising, Normalising, Yeo-Johnson Power Transformation, PCA Reduction.**

Standard Algorithms: **Linear Regression, Decision Tree Regressor, Lasso, Elastic Net, KNN Regressor, SVR**

Ensemble Algorithms: **Ada Boost Regressor, Gradient Boosting Regressor, Random Forest Regressor, Extra Trees Regressor**

Overall Result:

- *Overall Model of Choice - **Extra Trees Regressor** (1000 estimators) normalised + power transform + PCA (95%)*
- *Mean R2 (across 10 folds) - **0.792***
- *Standard Deviation (across 10 folds) - **0.024***

Conclusion and Future Works

- Focusing on **better feature engineering** will bring out better results with possibly the use of **simpler algorithms**
- Change of metrics could bring about different result (e.g. MSE, RMSE)
- Implement **grid search, randomised search or Bayesian optimisation** across **pipeline**
- Test other feature extraction methods
  - ***Singular Value Decomposition***
  - ***Iso-map Embedding***
  - ***Locally Linear Embedding***
  - ***Modified Locally Linear Embedding***
- **Discretisation** of **continuous** features to make them categorical
- Different **data split** methods
  - ***LOOCV***
  - ***Nested Cross Validation***
  - ***Repeated Random Train-Test Splits***
- Different **imputation** methods
  - ***Simple Imputer***
  - ***KNN Imputation***