A CUSTOMER SEGMENTATION STUDY ON AN E-COMMERCE HOUSEHOLD-GOOD STORE

EVERYTHING PLUS:

PLUS A LITTLE BIT MORE

by Jonathan Chan

Contents

1. Problem Formulation
2. Preliminary Hypotheses
3. Exploratory Data Analysis
4. Customer Segmentation – RFM
5. Customer Segmentation – K-Means Clustering
6. Findings & Conclusions

Section 1: Problem Formulation

- **Problem statement:**
  the client has specified a need for identifying distinct customer profiles out of their web channel – customer segmentation problem.

- **Primary objective:**
  to run personalized offers for distinct customer profiles.

- **Success indicator(s):**
  increase in conversion rates – running personalized offers will hopefully lead to greater user experience, of which will generate better brand salience and retention, while also aiding in word of mouth traffic. Target conversion rates: 5 – 7%

- **Time period:**
  approximately a year, from December, 2018 – November, 2019 (incl.)

- **Gap analysis:**
  qualitative research has been done, but none with data manipulation. The client does not possess intrinsic personal data on its customers; they only have transactional data at their disposal.

- **Key end user:**
  our analysis will be used by the client's product manager.

- **Key decisions:**
  unique personalized offerings will be made for each distinct customer profile we can identify.

Section 2: Preliminary Hypotheses

1. **Seasonality factors** – Sales are highest towards the end of the year.

2. **Day of the week transactions** – Customers are more likely to make purchases over the weekend than weekdays.

3. **Time elapsed between transactions** – Customers with shorter time deltas between transactions are more likely to exhibit higher repeat purchase behavior

4. **Purchase frequency (first order month) and average purchase frequency (proceeding months) are likely dependent on each other** – customers are likely to exhibit similar purchasing frequency in their lifetime based on their behavior in the first order month.

5. **Purchase frequency has a higher impact on LTV compared to average purchase value and average basket size** – customers who transact frequently are likely to generate higher lifetime value and subsequently, better loyalty and retention rates. This is irrespective of invoice value of that particular purchase or their quantities purchased.

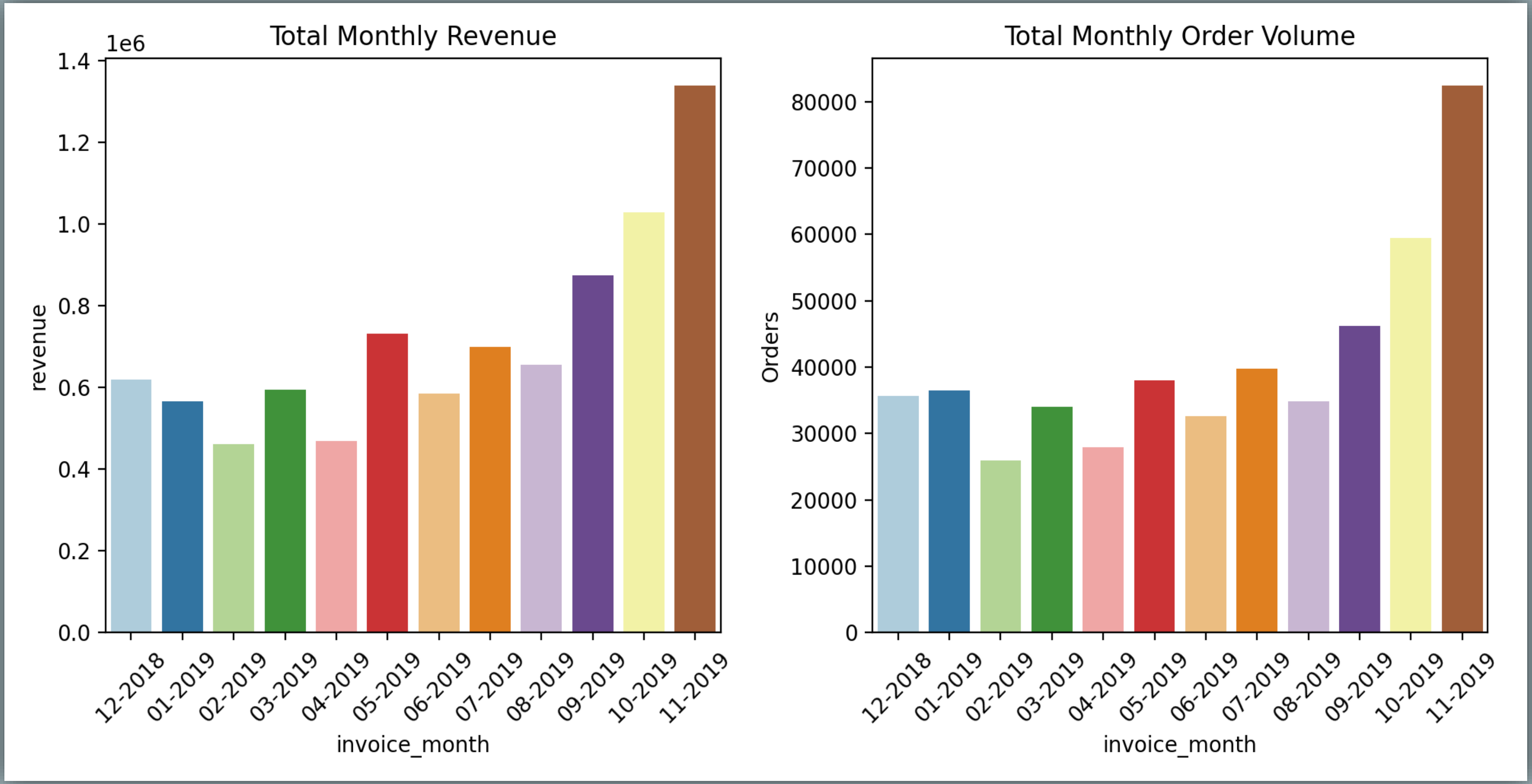
6. **Customer segments differ in terms of their average basket size (average quantity) and average frequency of purchase** – for example, one segment may consist of bulk buyers who purchase in larger quantities, while another segment may consist of more occasional shoppers with smaller purchase quantities.

Section 3:
Exploratory Data Analysis

SECTION NOTES:

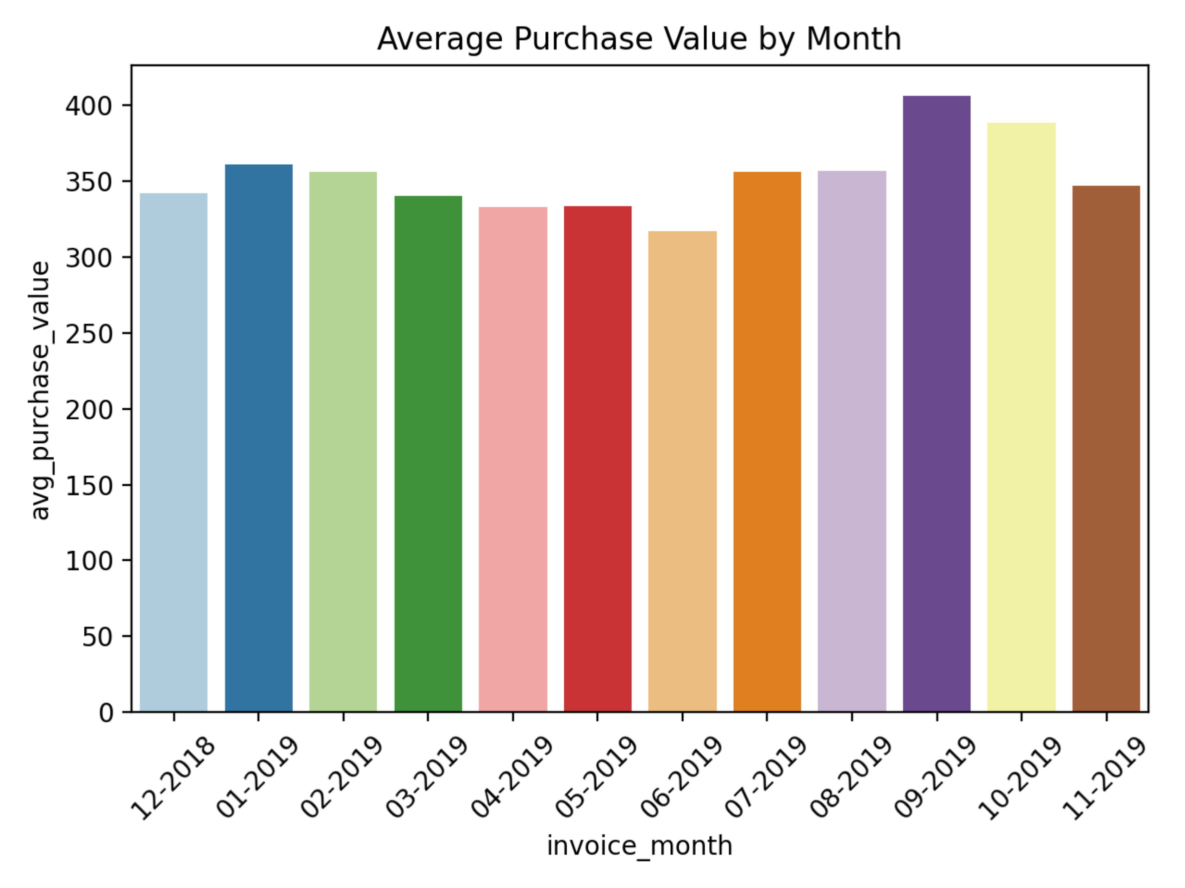
The following section aims to explore multi-variate feature relationships with the aim of testing our preliminary hypotheses. This will give us a fair amount of insight into the kind of customer behavior our dataset exhibits. In addition, this allows us to select only viable features/metrics to engineer for the purpose of customer segmentation modelling.

3.1. Seasonal Purchase Trends



- Total revenue and purchase volume increase dramatically towards the latter months of the year from **September** to **November**.
- **December** saw lower metrics all around, but we can attribute this to a lack of demand for household goods and an increase in demand for items of the 'gifting' nature, as we would assume is for the Christmas period.
- **September** marks the end of summer and the start of the schooling semester.
- **October** sees the Halloween festival come into play
- **November** ushers in thanksgiving and Black Friday.

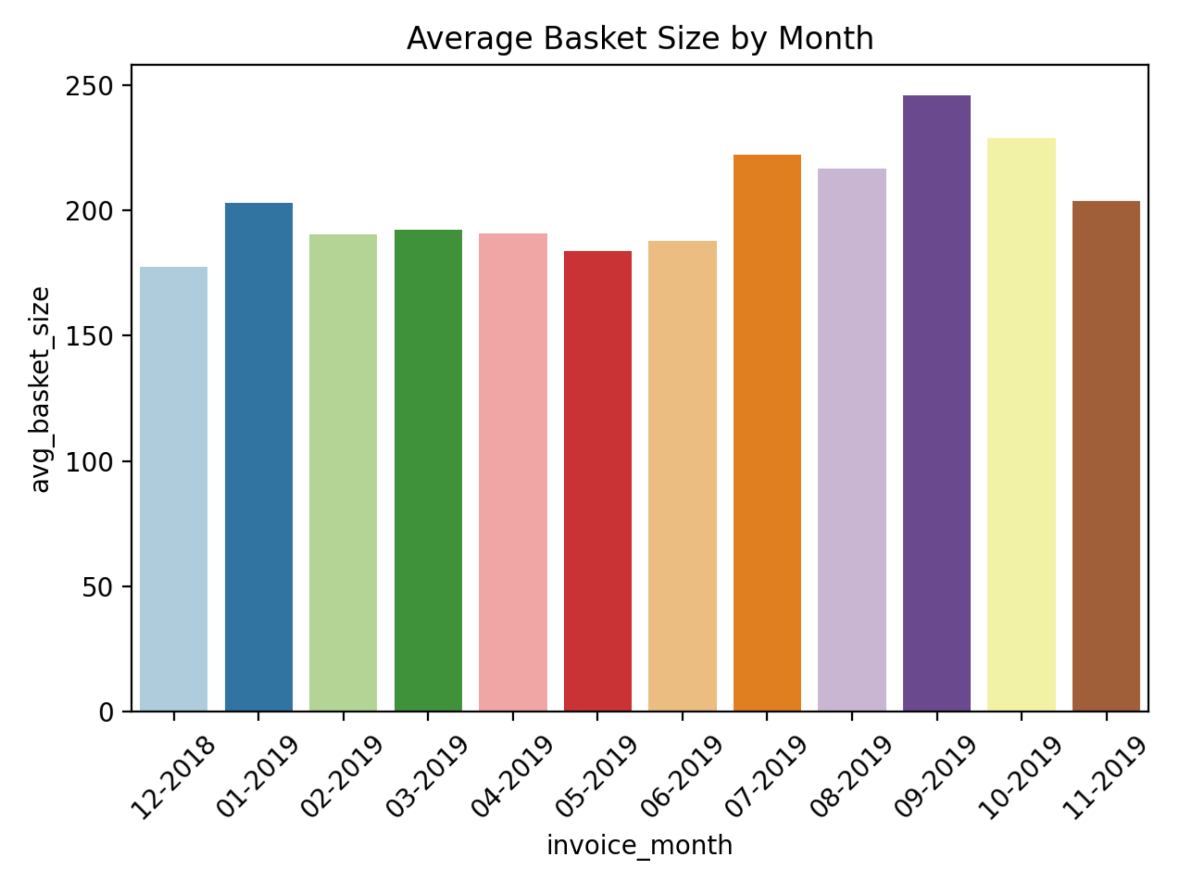
3.1. Seasonal Purchase Trends *cont.*



- Most high value transactions happen in the September and October. This possibly coincides with the end of Summer holidays and preparation for the festivities that occur from November onwards.

- The differences observed in Average Purchase Value by months aren’t as nuanced as expected. One thing to note is that the nature of the household good business is largely not of the ‘gifting’ type, which explains the lower purchase value values in festive months like November and December.

3.1. Seasonal Purchase Trends *cont.*



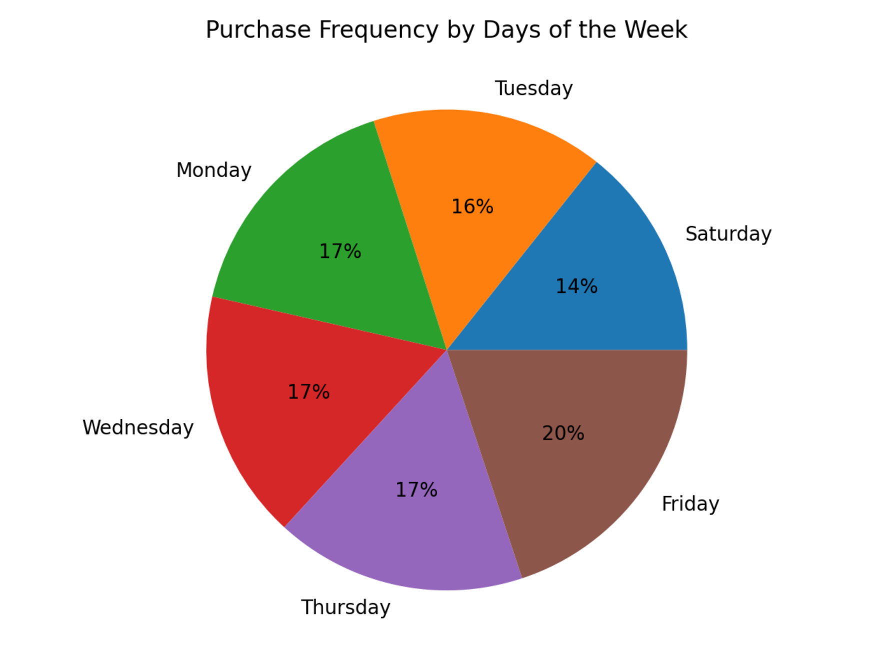
- A caveat about the nature of the household goods business: December being the month for Christmas, typically sees the purchase of gifts; we expect some form of transaction volume decrease for household goods. January sees big increase in basket size before continuously dipping towards the middle of the month in June.

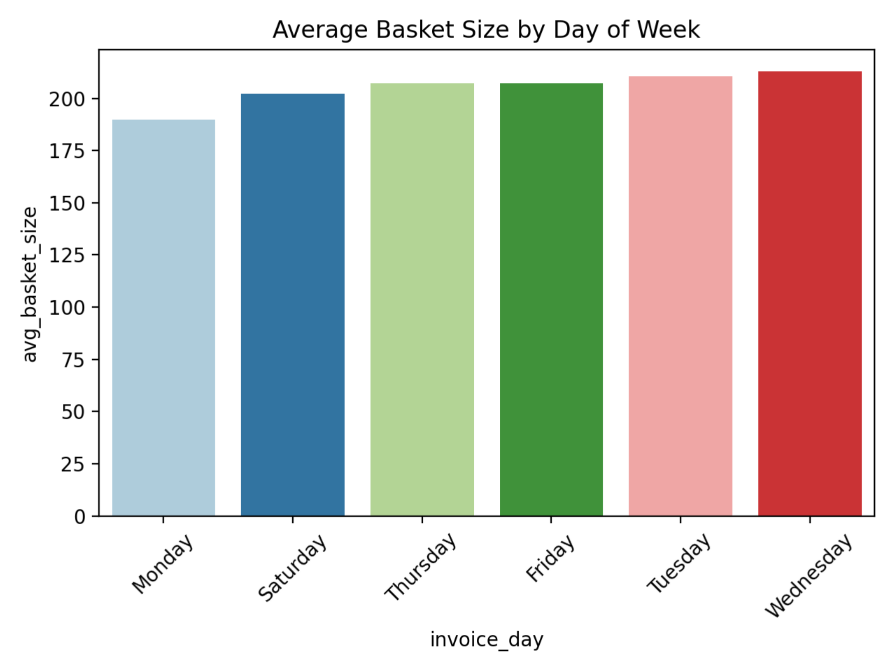
- We can attribute the high basket size values in January to be linked to post-Christmas clearances and the need to restock on household goods. As for the latter months on record, we yet again see increases in average basket size values during the start of summer and end of summer months, coupled with relatively higher values for November, which has Thanksgiving and Black Friday sales to thank for.

Hypothesis: Seasonal factors are present and sales are highest towards the end of the year is *Accepted*.

3.2. Purchase Freq. & Basket Size By Day Of Week

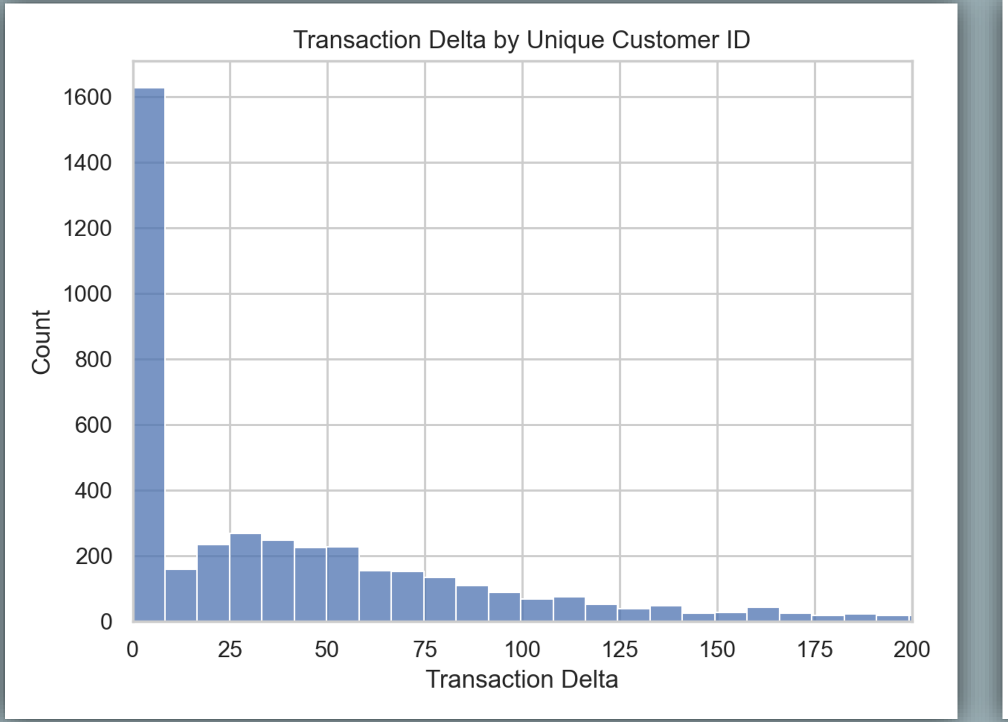
- **Note**: there are no records for Sunday. This could probably be explained by the e-commerce store going on break, or the store choosing not to record transactions on Sundays due to courier services going on break.
- Purchase Frequency is highest on **Friday**, while **Saturday** is the worst performing day.
- Average Basket Size is highest on Wednesday's, though the distribution across all weekdays are marginal at best.
- As we don't have a full weekend to work with, we cannot segment by purchase behaviors according to the days of the week.

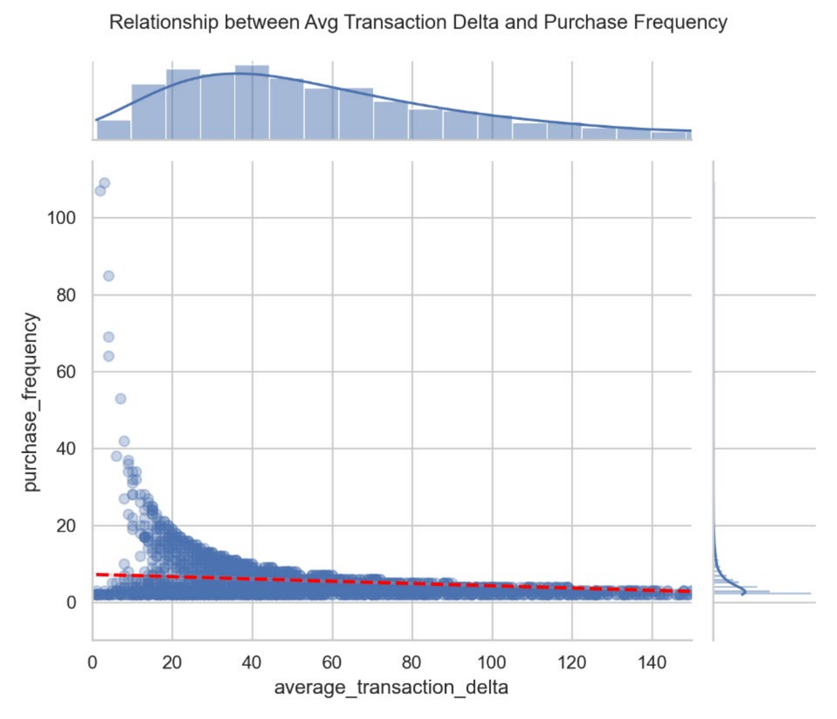




Hypothesis: Customers are more likely to make purchases over the weekend than weekdays is *Rejected*.

3.3. Transaction Time Delta & Total Purchase Frequency





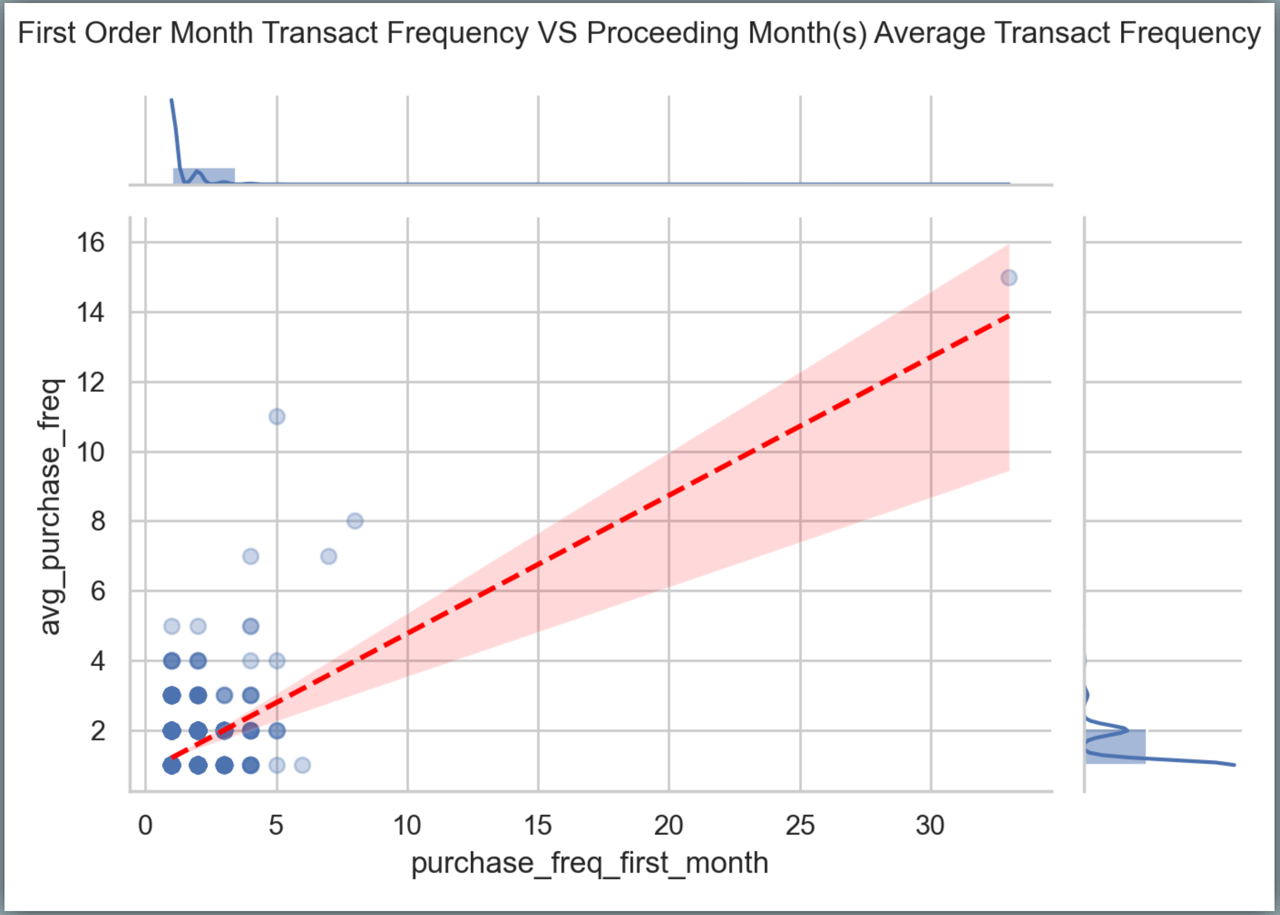
Amongst users with at least 2 lifetime transactions:

- **Spearman's Correlation Coefficient of -0.44976,** which indicates moderate negative correlation between a customer's average transaction delta and purchase frequency.
- When viewed from the right, correlation is somewhat stagnant before increasing in density starting from approximately the 90-day to 2-day time delta.

- The majority of customers belong to the 0-day time delta group, which represents "one-and-done" customers, or customers who made all their transactions within a 12-hour window.

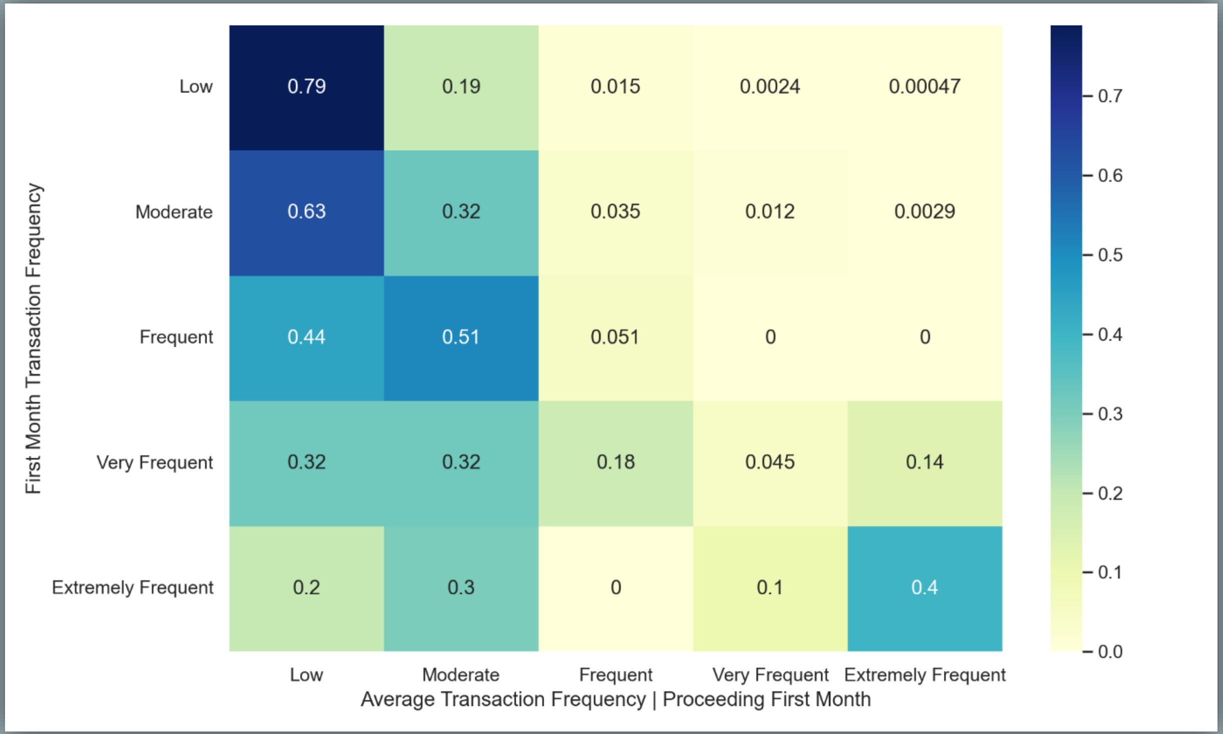
Hypothesis: Customers with shorter time deltas between transactions are more likely to exhibit higher repeat purchase behavior is *Accepted*.

3.4. First-month Purchase Frequency & Average Proceeding Lifetime Purchase Frequency



- Our joint-plot exhibits a very scattered array of data points, showing near to no apparent correlation.
- **Spearman's Correlation Coefficient of -0.2,** which indicates very low negative correlation between a customer's first month purchase frequency and their proceeding monthly frequency.
- We could go as far as to deduce that both variables are near independent of each other.

3.4. First-month Purchase Frequency & Average Proceeding Lifetime Purchase Frequency *cont.*



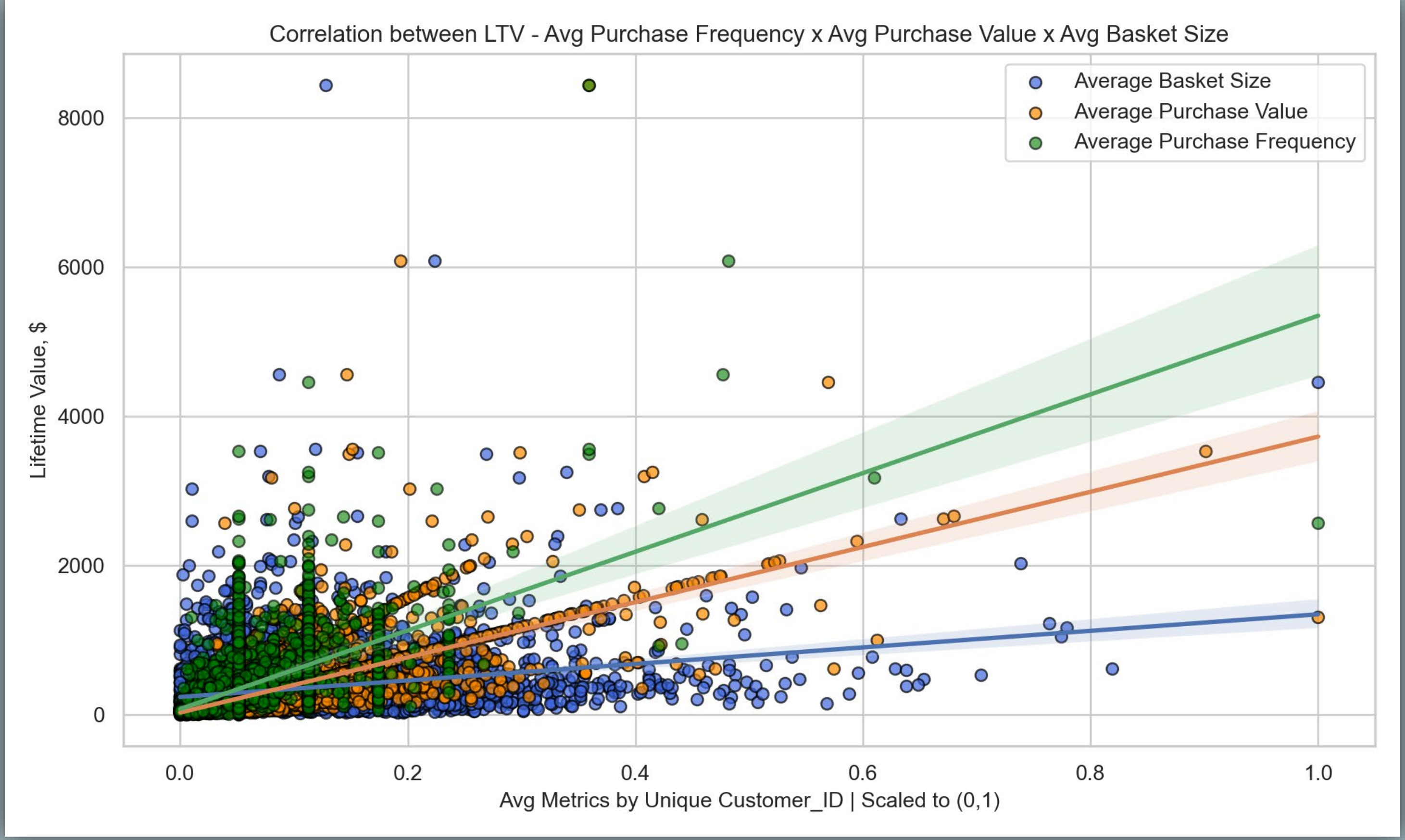
First Month Transaction Frequency	Low	Moderate	Frequent	Very Frequent	Extremely Frequent
Low	0.79	0.19	0.015	0.0024	0.00047
Moderate	0.63	0.32	0.035	0.012	0.0029
Frequent	0.44	0.51	0.051	0	0
Very Frequent	0.32	0.32	0.18	0.045	0.14
Extremely Frequent	0.2	0.3	0	0.1	0.4

A Chi-test of Independence was performed:

- Our contingency table, to prove our Alternate Hypothesis (dependence between both variables) true, we'd like to see more deep blue cells across the middle diagonal. Unfortunately we did not, and the only dependent frequency group are those that purchased once a month in the first order-month and subsequently made an average of 1 purchase across their subsequent lifetime.
- Our Chi-Test's **p-value of 99.9%** also proved this theory to be true, as assuming a 5% level of confidence, we have no reason to reject our Null hypothesis. **Both variables are deemed independent of one another.**

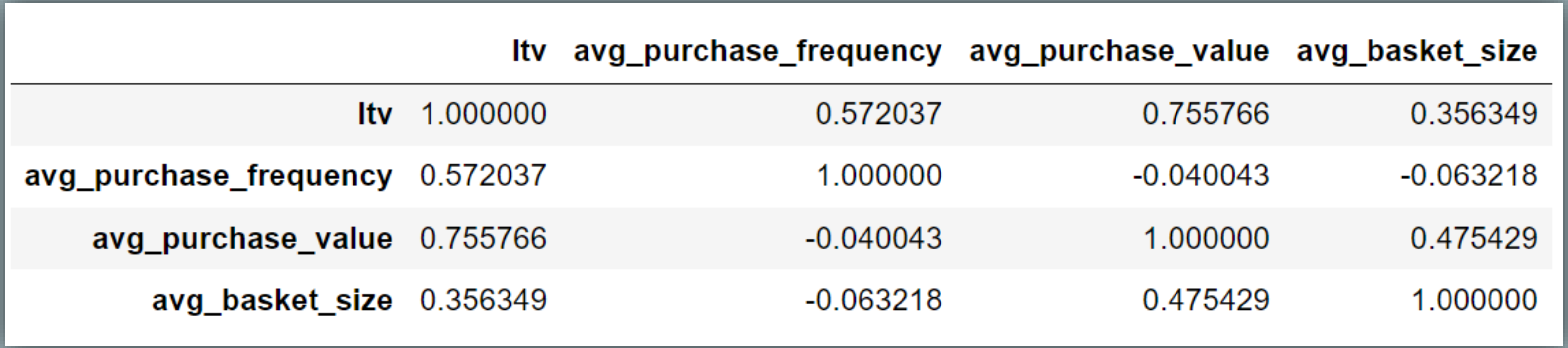
Hypothesis: Purchase frequency (first order month) and average purchase frequency (proceeding months) are likely dependent on each other – customers are likely to exhibit similar purchasing frequency in their lifetime based on their behavior in the first order month is *Rejected*.

3.5. Relationship Between Ltv & (Purchase Frequency | Purchase Value | Basket Size)



- Amongst our three metrics against Customer Lifetime Value, **Average Purchase Value** achieved the best positive correlation-coefficient score of **0.76**, followed closely by **Average Purchase Frequency** at **0.57**, and lastly **Average Basket Size** with a low correlation of **0.36**; this is depicted in our scatterplot, whereby our regression line-plots show the approximate direction of our metrics.
- Hence we will infer that both purchase frequency and purchase value have equal importance in both present and future LTV indication, with the latter taking precedence over the other.

3.5. Relationship Between Ltv X (Purchase Frequency | Purchase Value | Basket Size) *cont.*



	ltv	avg_purchase_frequency	avg_purchase_value	avg_basket_size
ltv	1.000000	0.572037	0.755766	0.356349
avg_purchase_frequency	0.572037	1.000000	-0.040043	-0.063218
avg_purchase_value	0.755766	-0.040043	1.000000	0.475429
avg_basket_size	0.356349	-0.063218	0.475429	1.000000

Spearman’s correlation matrix:

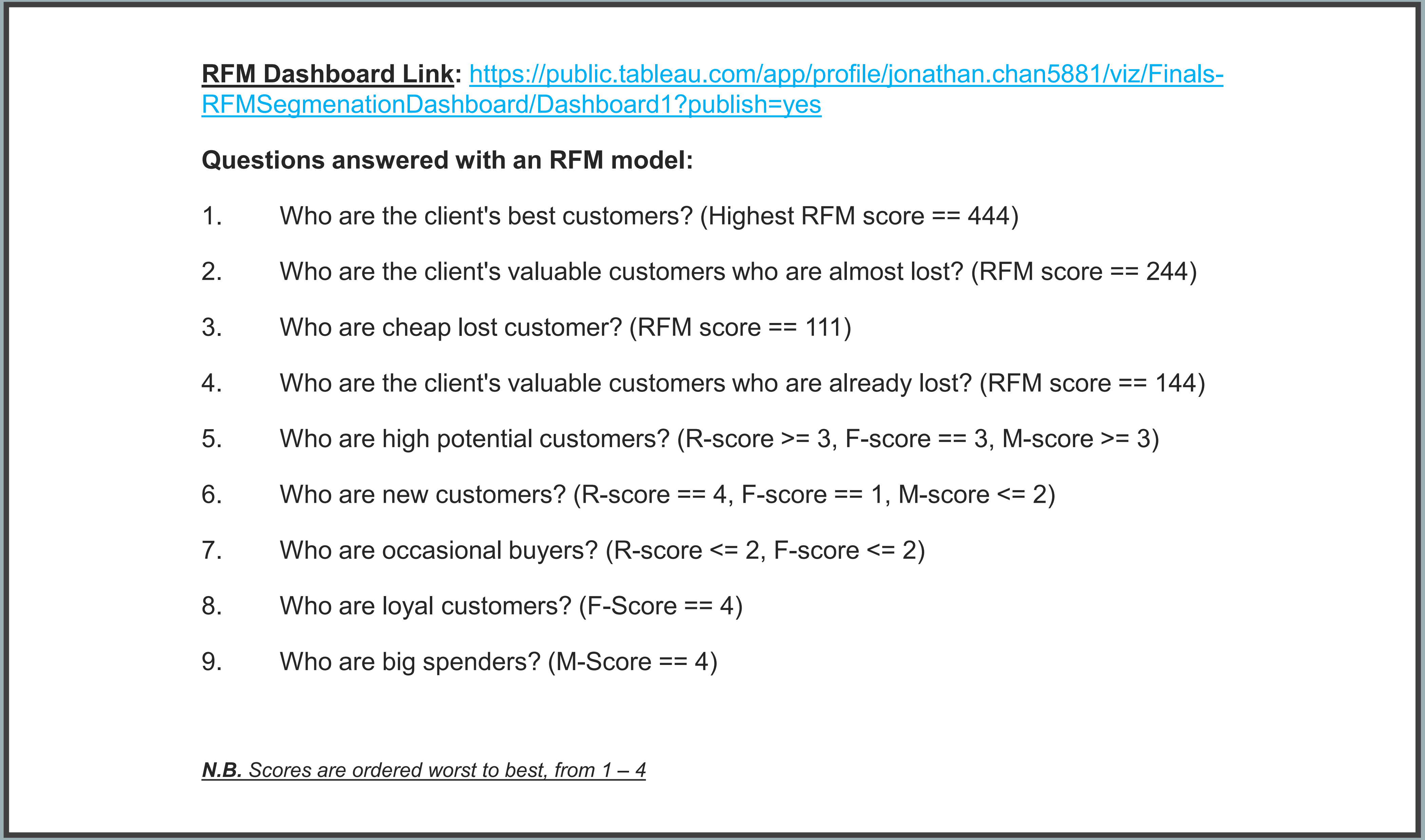
- Observing the first row, we can explain the correlations between our metric features and LTV variable numerically. **Average Purchase Value is the most dominant metric in relationship with LTV**.
- We can also discern correlations between metrics within themselves. One that stands out is the **close to Null correlation between Average Basket Size and Average Purchase Frequency**

Hypothesis: (1) Purchase frequency has a higher impact on LTV compared to average purchase value and average basket size & (2) Customer segments differ in terms of their average basket size (average quantity) and average frequency of purchase are both *Rejected*.

Section 4:
Recency – Frequency – Monetary Segmentation Model (RFM)

SECTION NOTES:

The RFM model comprises tabular data of RFM scores assigned to each unique customer ID in our dataset. The scores for each feature of Recency, Frequency and Monetary are calculated based off their distinct quartile range of values (<0.25%, <0.5%, <0.75%, <1.0%), and are given weightage classes from 1 – 4, where 1 is worst and 4 is best. A ready-made Tableau dashboard will be used to create customer segments, and helps the client perform more granular analysis.



RFM Dashboard Link: https://public.tableau.com/app/profile/jonathan.chan5881/viz/Finals-RFMSegmenationDashboard/Dashboard1?publish=yes

Questions answered with an RFM model:

1. Who are the client's best customers? (Highest RFM score == 444)
2. Who are the client's valuable customers who are almost lost? (RFM score == 244)
3. Who are cheap lost customer? (RFM score == 111)
4. Who are the client's valuable customers who are already lost? (RFM score == 144)
5. Who are high potential customers? (R-score >= 3, F-score == 3, M-score >= 3)
6. Who are new customers? (R-score == 4, F-score == 1, M-score <= 2)
7. Who are occasional buyers? (R-score <= 2, F-score <= 2)
8. Who are loyal customers? (F-Score == 4)
9. Who are big spenders? (M-Score == 4)

***N.B.** Scores are ordered worst to best, from 1 – 4*

Section 5:
K-Means Segmentation Model

SECTION NOTES:

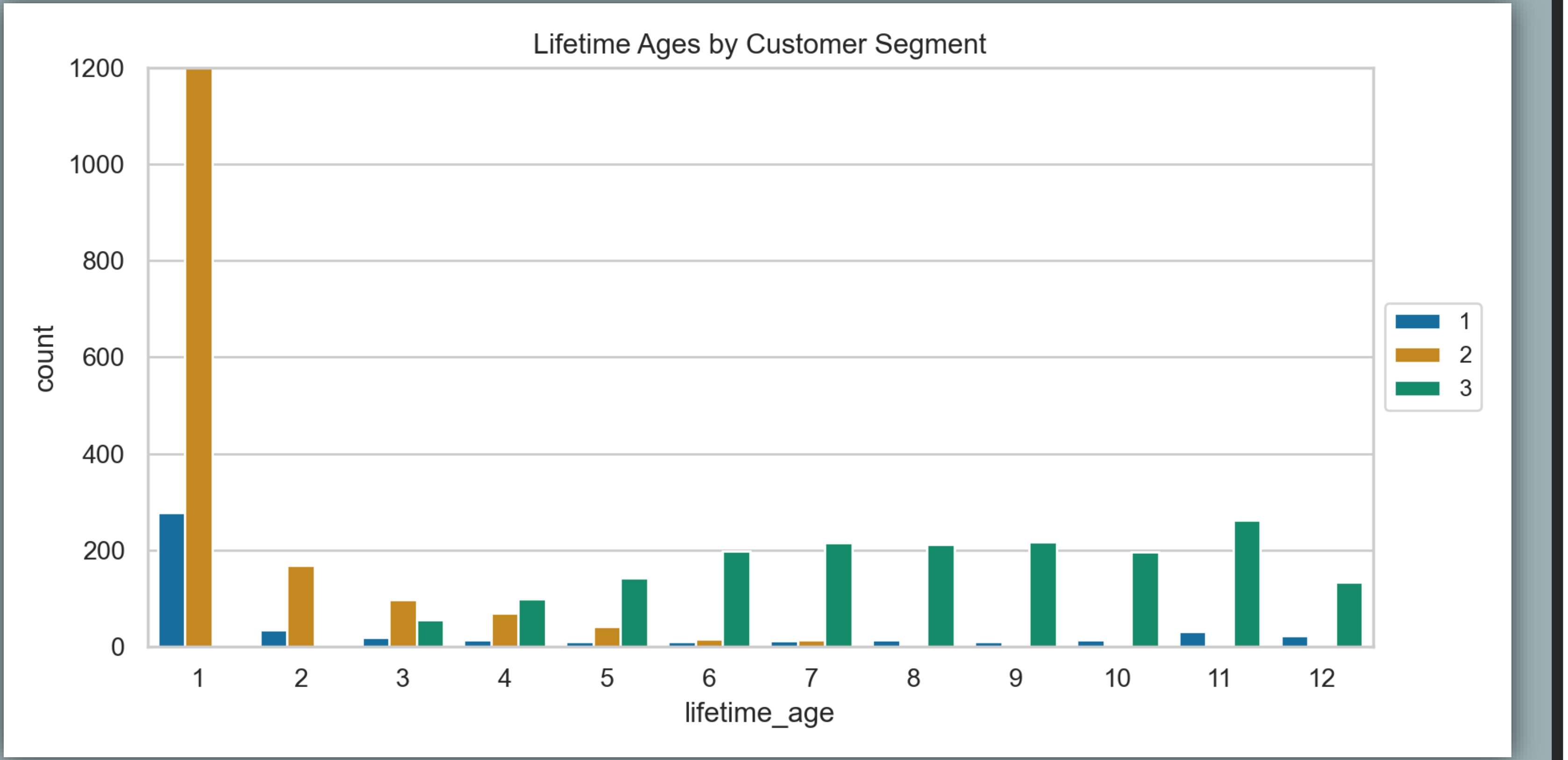
The K-Means model uses a machine learning clustering algorithm to segment users based on their similarities to one another based on a matrix of features/metrics. We clustered our dataset's customers into 3 segments as this achieved the best clustering evaluation scores (silhouette and inertia). **Segments are numbered 1 – 3, sorted by highest Lifetime Value (LTV) to lowest.**

N.B. *LTV values (Segment 1 - $1123, Segment 2 - $289, Segment 3 - $212)*

The features we engineered for this purpose are as follows:

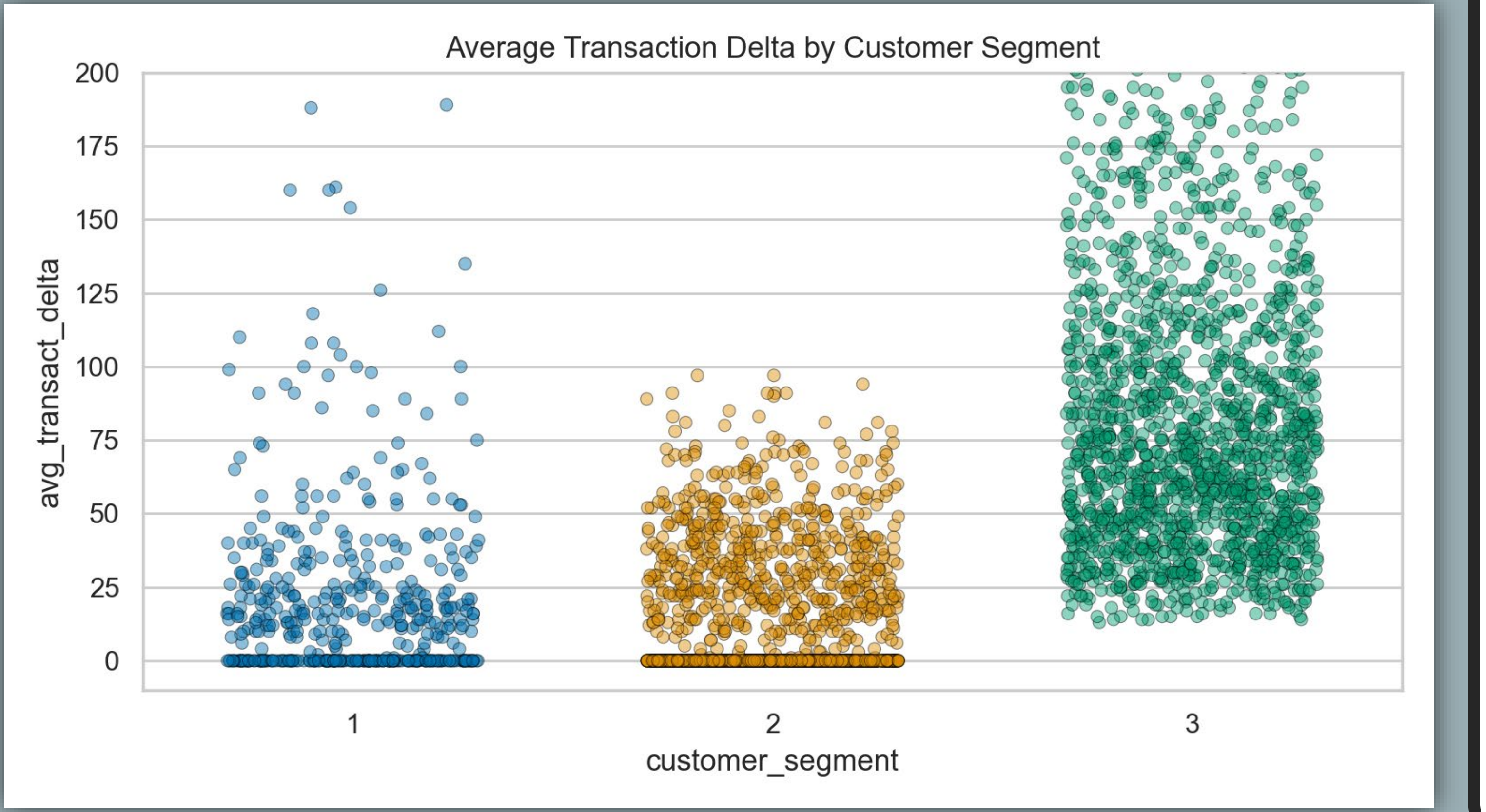
- Lifetime Age
- Price Sensitivity
- Seasonal Shoppers
- Average Transaction Delta
- Monthly Purchase Frequency
- Average Purchase Value
- Average Basket Size
- Lifetime Value (LTV)

5.1. Lifetime Age



- Firstly, note the smaller amount of observations for **Segment 1**. These customers are those with relatively larger values for LTV, and assume the title of bulk/large basket size purchasers by default. Their lifetime values are concentrated in the 1-month age, but are consistently seen across all lifetime month intervals. (average – 3.42 months)
- Majority of **segments 2** customers fall into the 1-month lifetime group. This group declines rapidly from 2 months onward (average – 1.5 months).
- Lastly, **segment 3** has the largest lifetime values, although it should be noted their LTV is lowest amongst our 3 segments. This could represent customers who are periodic/occasional shoppers, but are loyal to the brand. (average – 8.12 months)

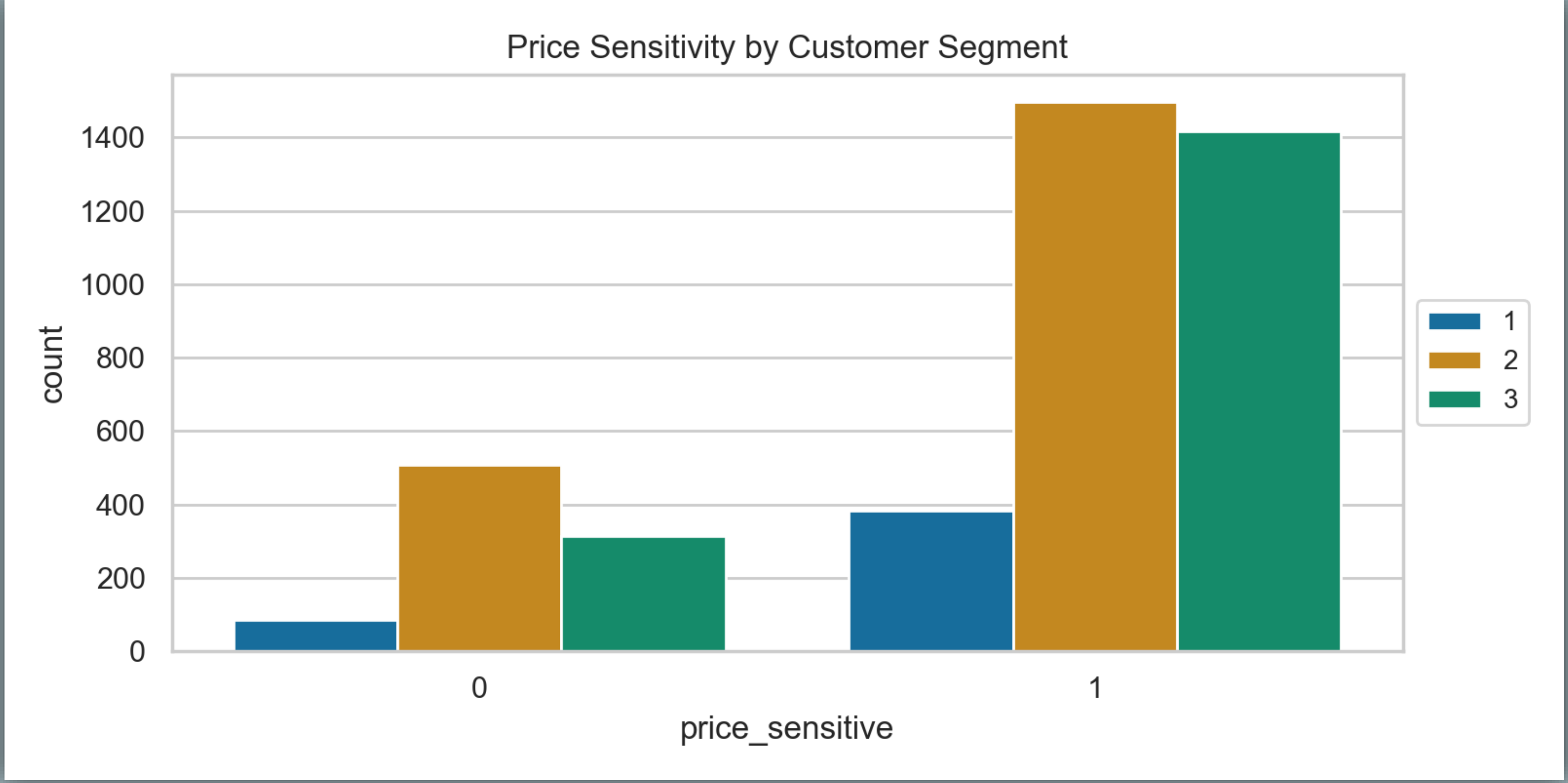
5.2. Average Transaction Delta



- **Segments 1** (21-day avg) **and 2** (10.7-day avg) contain large amounts of users who transact below a 1-day time delta. These are customers who are one-and-done shoppers, as observed in our Lifetime Age graph. Apart from this, notice the dispersion of values occurring in both groups which suggests the type of behavior that is seen in periodic shoppers. This observation is significantly more evident in **segment 1**, where we can see much more dispersion after the 50 day time delta. One caveat to note, is that **segment 1** has a relatively lower sample size compared to other segments. Overall, **segments 1 and 2** have significantly lower transaction time delta.

- **Segment 3** (94-day avg) has consistently larger average transaction deltas and extreme amounts of variation in its time delta observations. They do not however, have any one-and-done shoppers, and they could potentially be shoppers who are innately loyal to an extent, and looking at their mean delta values, could also represent customers who are periodic shoppers.

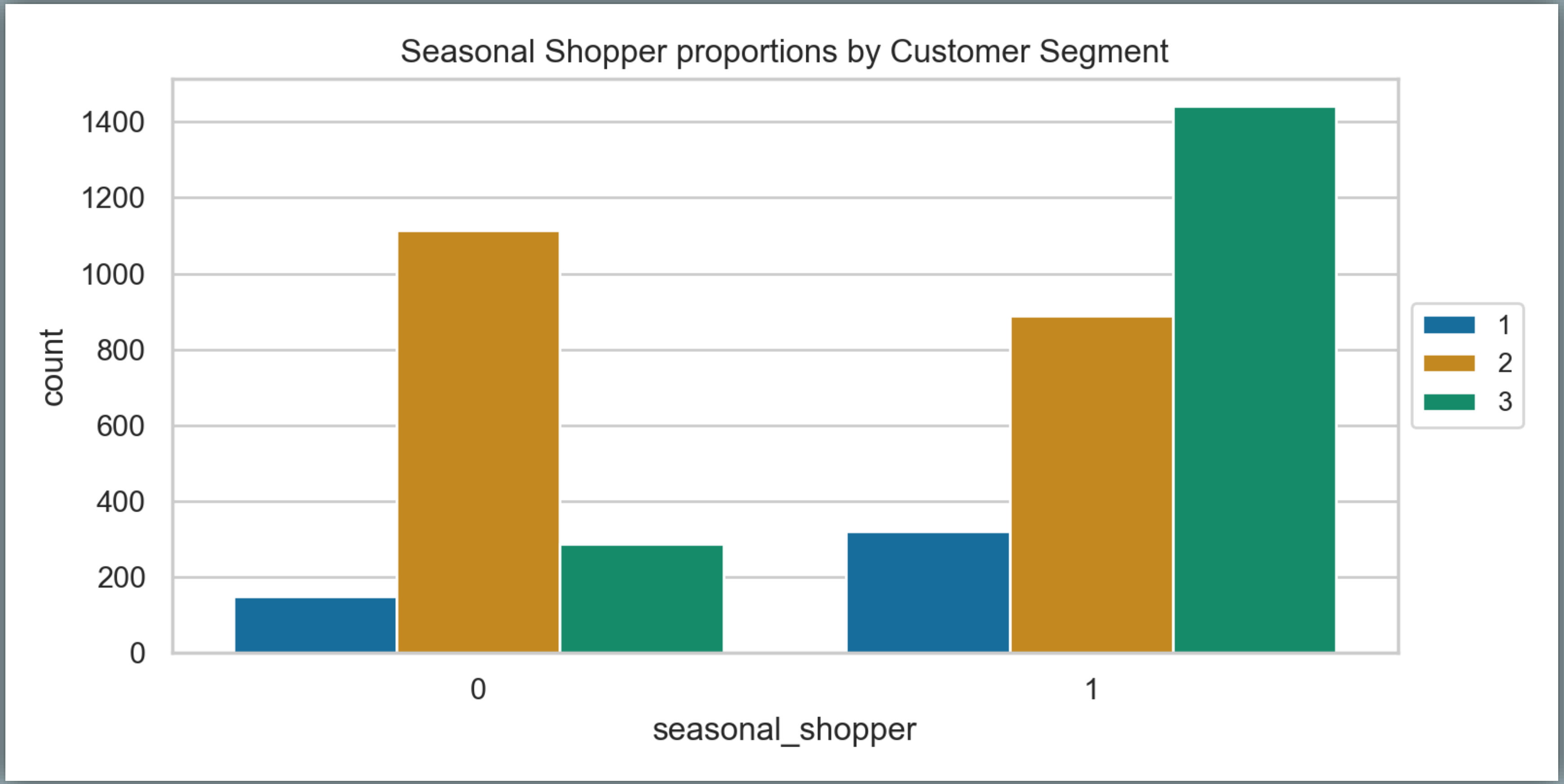
5.3. Price Sensitivity



- **Segment 2** is the least price sensitive comparatively (**75%** segment share). If we remove **Segment 1** (highest LTV but low sample size) and compare **segments 2 and 3** in a vacuum, the lower price sensitivity for **segment 2** checks out with its higher LTV.

- **Segment 1 and 3** (**82%** segment share) are innately similar in their price sensitivity, though there is a stark difference in their LTV values.

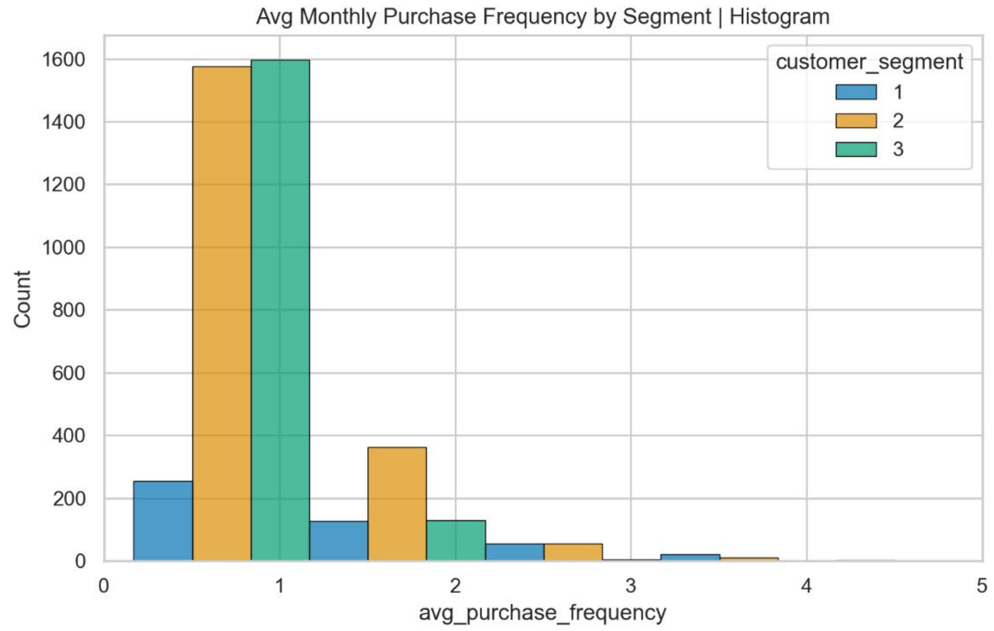
5.4. Seasonal Shoppers

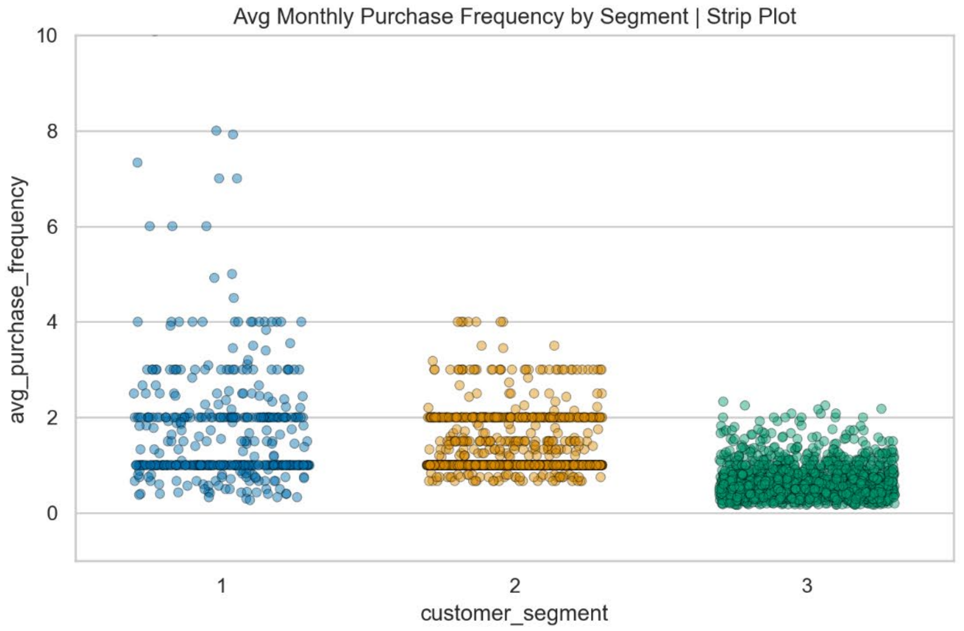


- **Segment 3** comprises customers who rely on seasonal bargains for their shopping needs at a **83%** seasonal shopper share, followed closely by **segment 1** at a **68%** share.

- **Segment 2** stands out to us, as its customers have the least seasonal shopper proportion at a 44% share. This ties in with our price sensitivity graph, which showed **segment 2** being the least price sensitive amongst all segments.

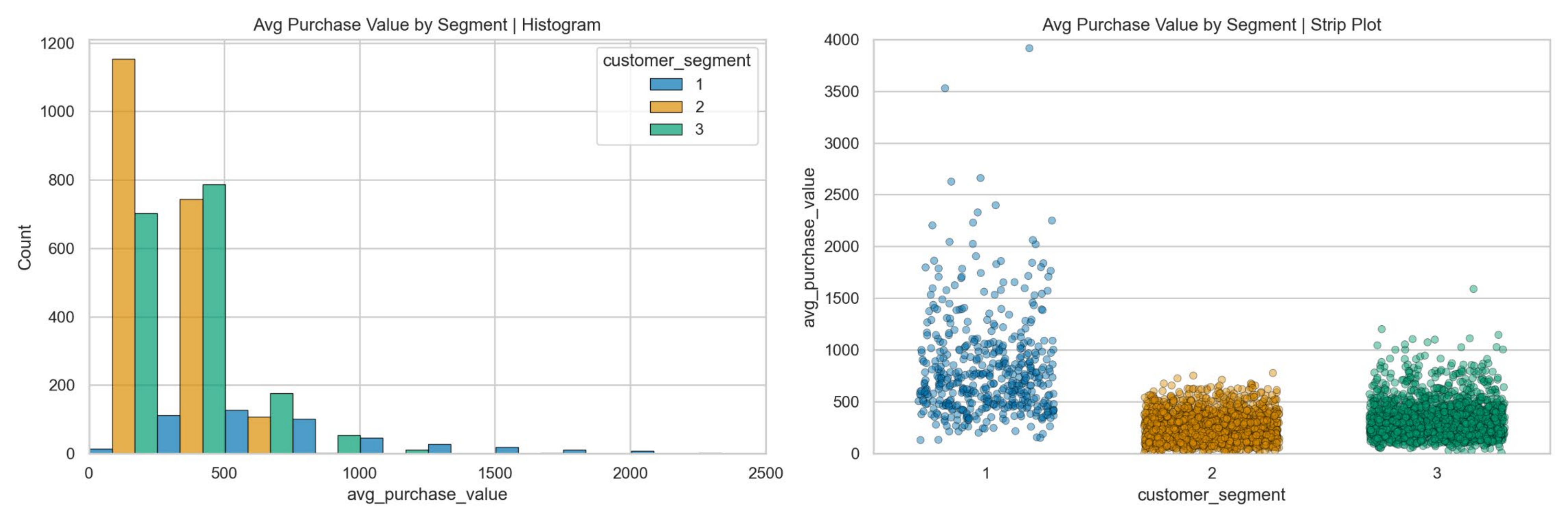
5.5. Monthly Purchase Frequency


Avg Monthly Purchase Frequency by Segment | Histogram


Avg Monthly Purchase Frequency by Segment | Strip Plot

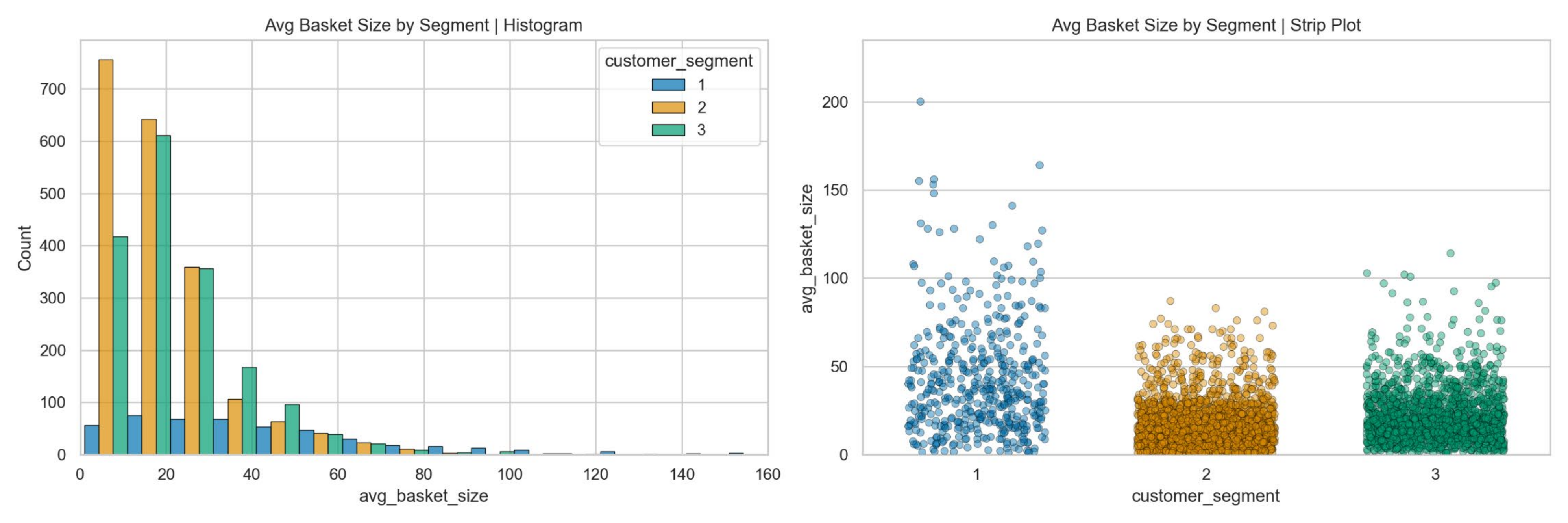
- Average purchase frequency for **segment 1** **and 2** are roughly similar in terms of median values with 1 purchase a month, though looking at their mean values, **segment 1** (1.69 avg) has higher variance towards the upper end compared to **segment 2** (1.20 avg) and this is succinctly depicted in our scatterplot.
- **Segment 3** (0.64 avg) as noted before, resembles customers who are periodic shoppers and this is once again seen here, averaging less than 1 transaction per month.
- Amongst our 2 segments with high sample sizes (2 and 3), **segment 2** leads both segments in the twice-a-month frequency group and beyond that, which is something noteworthy.

5.6. Average Purchase Value



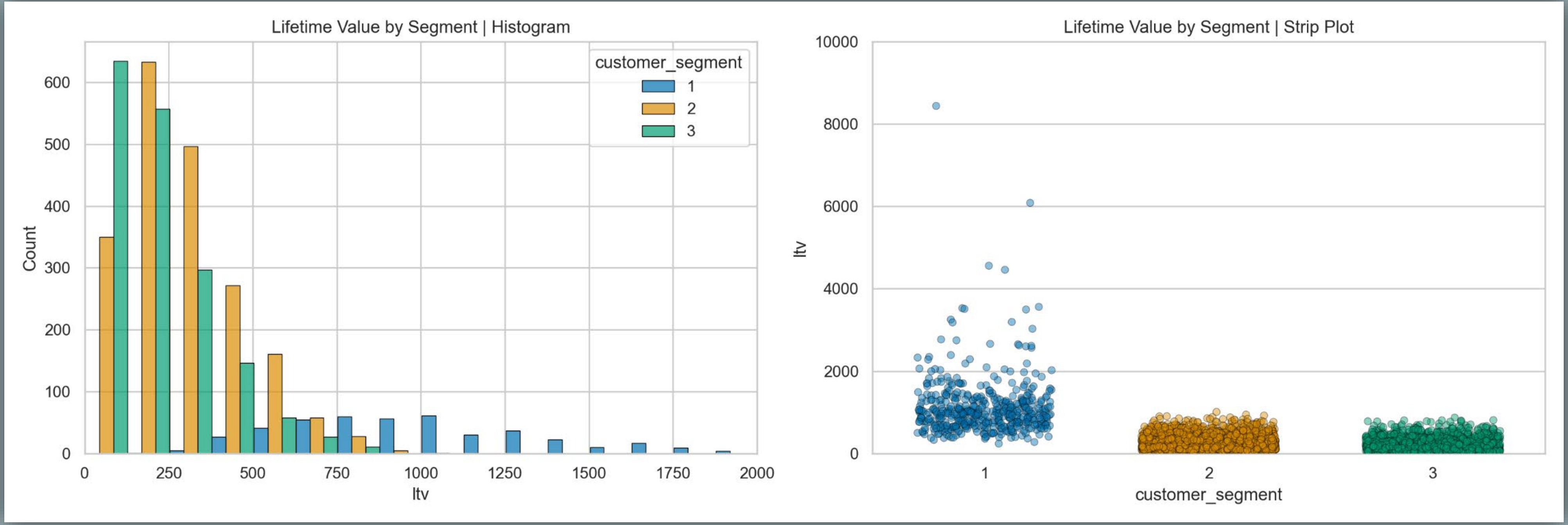
- **Segment 1** as usual has immense dispersion of distribution. Here though, we can confirm our prior assumption of its customers being bulk purchasers for purposes such as reselling, institutional purchases or corporate gift giving, as we observe their mean average purchase value of $821.
- **Segment 2's** average purchase value hovers around the $247 mark as represents the lowest observable value amongst our segments. The majority of **Segment 3's** customers are clustered at or below the $500 mark and has a mean average of about $331.
- Overall, if we recall the fact that **segment 2** had the lowest transaction delta on record, we could deduce that these customers are those that purchase smaller amounts more frequently, while the converse can be said about **segment 3**.

5.7. Average Basket Size



- Yet again, **segment 1** has high basket size variance, with a mean average of 42 and median average of 37. These customer typically purchase larger quantities compared to our other segments and tie in with our assumption about them being bulk shoppers.
- Our plots of average basket size confirms our suspicion about **segment 2 and 3**. In terms of relative proportions to other segments, **segment 2** typically purchases lower quantities (17 basket avg) at a higher frequency, while **segment 3** typically purchases bigger quantities (21 basket avg) at lower frequencies.

5.8. Lifetime Value (LTV)


Lifetime Value by Segment | Histogram

Lifetime Value by Segment | Strip Plot

- **Segment 1's** average LTV ($1123) is no doubt the highest amongst our bunch of segments, with much larger basket size averages and relatively higher purchase frequencies.
- **Segment 2's** LTV ($289) is middling of the road, and tops **segment 3** purely due to its relatively lower price sensitivity, lower transaction delta and higher purchase frequency. A caveat to note however, is that the lifetime age for this segment is extremely low, at about 1.5 months.
- **Segments 3's** LTV ($211) is lowest amongst all segments, and this can be attributed to its very sparse purchase frequency and large transaction time deltas, though it is important to note they comprise of loyal customers who have been with the business for extended period of time, or an average of 8 months.

Section 6: Findings & Conclusions

SECTION NOTES:

Our Findings & Conclusions section details the consolidated summary of our K-Means model, comprising our 3 segments and their defining behavioral attributes. As stated earlier, the client may use the RFM dashboard at its own discretion for a more granular approach to segmenting its customers, with the model's ability to filter for a multitude of customer behavioral characteristics. Otherwise, the K-Means segmentation results provides the client with a macro snapshot of 3 distinct segments which can be acted upon instantly from a marketing standpoint.

Segment 1

Behavioural Traits:

- Average of 3.43 lifetime months
- Average transaction time-delta of 21 days - Comprises a large cluster of one-and-done shoppers. Remaining customers are repeat purchasers who are either frequent purchasers within a month time-delta, or periodic shoppers with time deltas as far as a 6 months time-delta.
- Most price sensitive (82% share)
- 2nd largest segment share of seasonal shoppers (68% share)
- Average monthly purchase typically once-a-month, with variance towards the twice-a-month range and beyond (1.69 avg)
- Average basket size of about 42, which is highest amongst all segments
- Highest purchase value with a mean of $821
- Highest LTV, at a mean of $1123

Recommendations:

This segment comprises customers who are largely bulk buyers. This was inferred via its largest average basket size, with a substantial margin to the segment with the next largest basket size at 21. This could be for a multitude of reasons, like institutional purchases, reselling or corporate gifting. While segment 1 has a lower relative sample size compared to our others, their LTV values stand out to us and comprise a significant amount of revenue for the company. This is compounded by the fact this segment typically makes repeat purchases within a month or less, which indicates some loyalty in a vacuum.

Its low lifetime age of 3.43 months means we have not really seen the full extent of this group's behavior, and we cannot determine if there is innate loyalty to the brand until a later date. Couple this with the fact that this segment has the highest average basket size indicates it can be profitable if we can encourage them to continue their purchase behavior in the future. There is an opportunity for a loyalty program that rewards customers for repeat bulk purchases. One thing to note, is that these customers are those who enjoy a good bargain, with 82% share of them being price sensitive and a good amount of them being seasonal shoppers. The client may want to implement cross-selling recommendation blocks for similarly affordable goods, which may further increase repeat purchase behavior.

Segment 2

Behavioural Traits:

- Average of 1.5 lifetime months
- Average transaction time-delta of 10.7 days - Comprises a large cluster of one-and-done shoppers, a moderate cluster of frequent buyers within a 1 month time delta, and periodic shoppers no larger than 100-days time delta.
- Lowest price sensitivity (75% segment share)
- Smallest customer proportion of seasonal shoppers (44% share)
- Average monthly purchase typically once-a-month, with some variance towards twice-a-month. (1.2 avg)
- Lowest observed average basket size with a mean of 17
- Lowest observed average purchase value, with a mean of $247
- 2nd-highest LTV, at a mean of $289

Recommendations:

This segment has great potential but needs to be monitored. Its lowest lifetime age indicates we have yet to see the full extent of the segment's behavior and relatedly, their loyalty. While this segment exhibited the lowest metrics for average basket size and average purchase value, there are some key highlights, which would be its higher propensity to purchase more frequently, its lowest proportion share of seasonal shoppers and relatedly, its lowest price sensitivity. This accordingly has resulted in the 2nd highest LTV amongst our segments, possibly due to its higher transaction frequency and a somewhat disinclined attitude towards price consciousness.

The segment's higher purchase frequency behavior, coupled with lower transaction-delta and disinclination towards bargains and cheaper goods (to an extent) can be leveraged to increase overall customer value and loyalty. While it needs more monitoring, potential is there to upsell slightly more premium product ranges. In terms of improving this segment's low basket size and purchase value, the client may want to explore cross-category promotions that incentivizes these customers into exploring different product lines.

Segment 3

Behavioural Traits

- Average of 8.1 lifetime months
- Average transaction time-delta of 94 days - No one-and-done shoppers are observed. Comprises loyal but periodic shoppers, as average transaction delta is the highest observed amongst segments. Majority of observations are seen scattered below the 100-day time delta, while a significant portion of other observations are seen above that.
- Most price sensitive (82% share)
- Largest amount of seasonal shopper proportion (83% share)
- Lowest average monthly purchase frequency at roughly 0.64/month average.
- 2nd-highest average basket size, with a mean average of 21
- 2nd-highest observed average purchase value, with a mean average of $331
- Lowest LTV, at a mean of $211

Recommendations:
This segment of customers are our most loyal/oldest customers as they average approximately 8 months of lifetime age values. The biggest drawbacks are the fact that they do not purchase as frequently as other segments, with the largest observed transaction delta and lowest average purchase frequency on record. They also represent the segment with the highest share of seasonal shoppers and are solely focused on price consciousness. As they are our most loyal customer base, the biggest benefit from this is the potential for word of mouth traffic and brand advocacy.
This segment should be given the most attention, as it comprises our longest and most loyal customers. Its long lifetime age also indicates we are more or less sure of its behavioral characteristics and can act without any assumptions and biases. The main notion is that with brand loyalty, comes increased brand salience. The client may want to create referral programs and(or) incentives for customer advocacy that will increase purchase frequency as well as generate new-user acquisition.