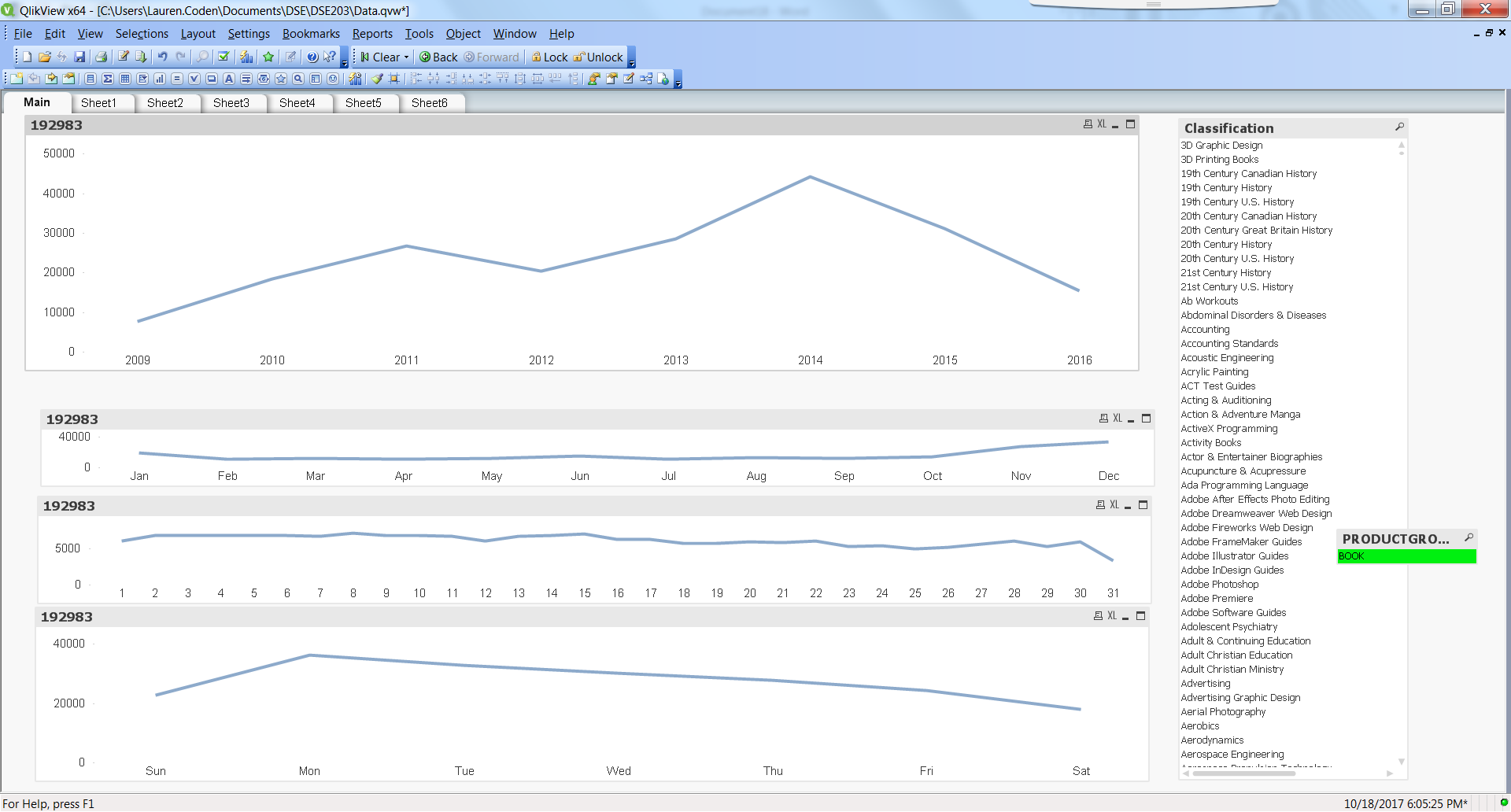
**User Visualizations (research demographic/category groupings)**

Research possible useful features for collaborative learning (creating and finding similar user profiles) via visualizations/metrics (distributions, scatterplots, clustering, correlations). Identifying how the features are distributed over all the orders would be a good starting point. Initial thought is that features that have some variance and some non-random grouping are more useful for similarity (cluster) based models.

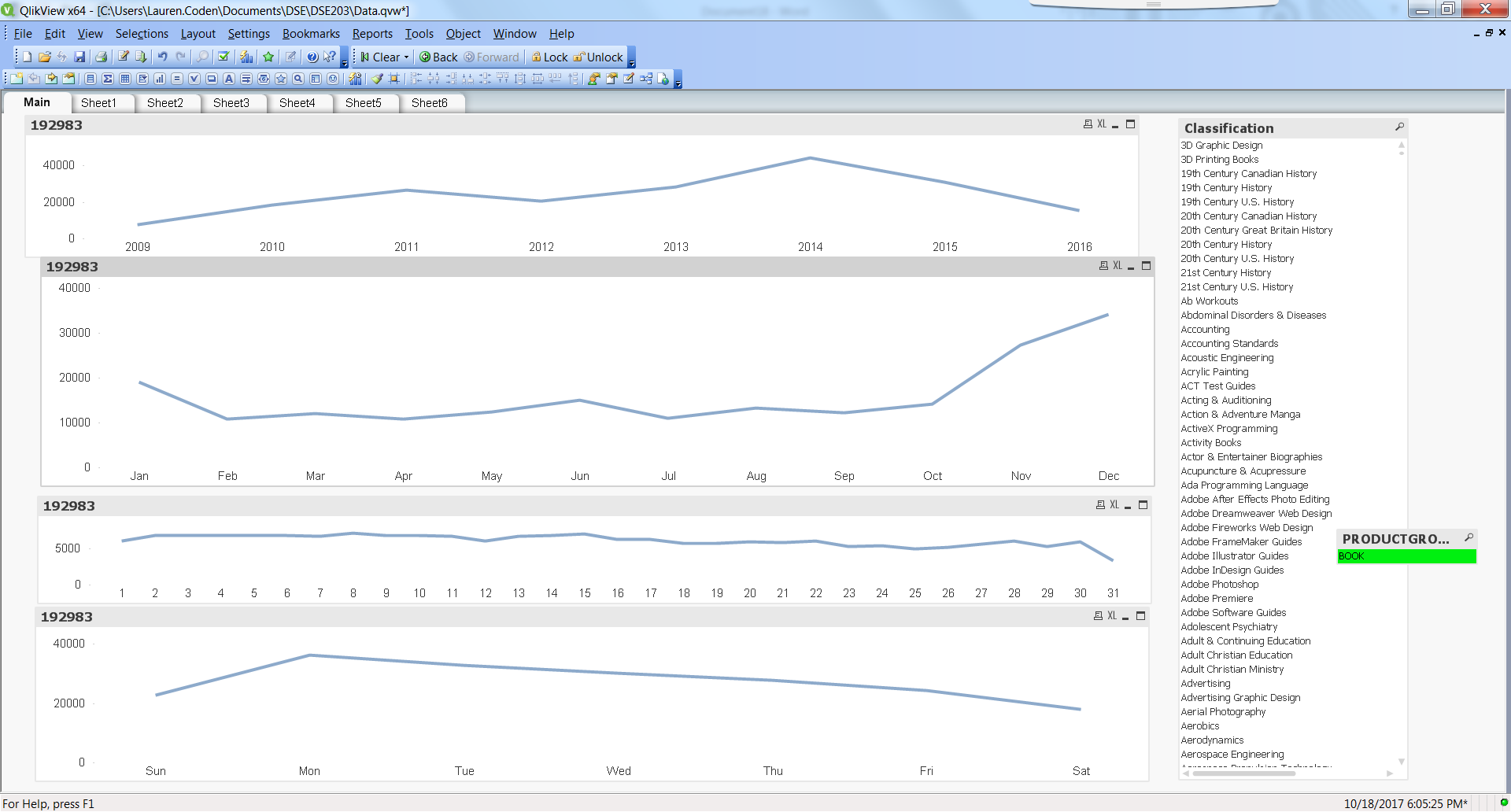
**Distribution visuals/Observations for possible buckets (eyeballing):**

**Time Buckets:**

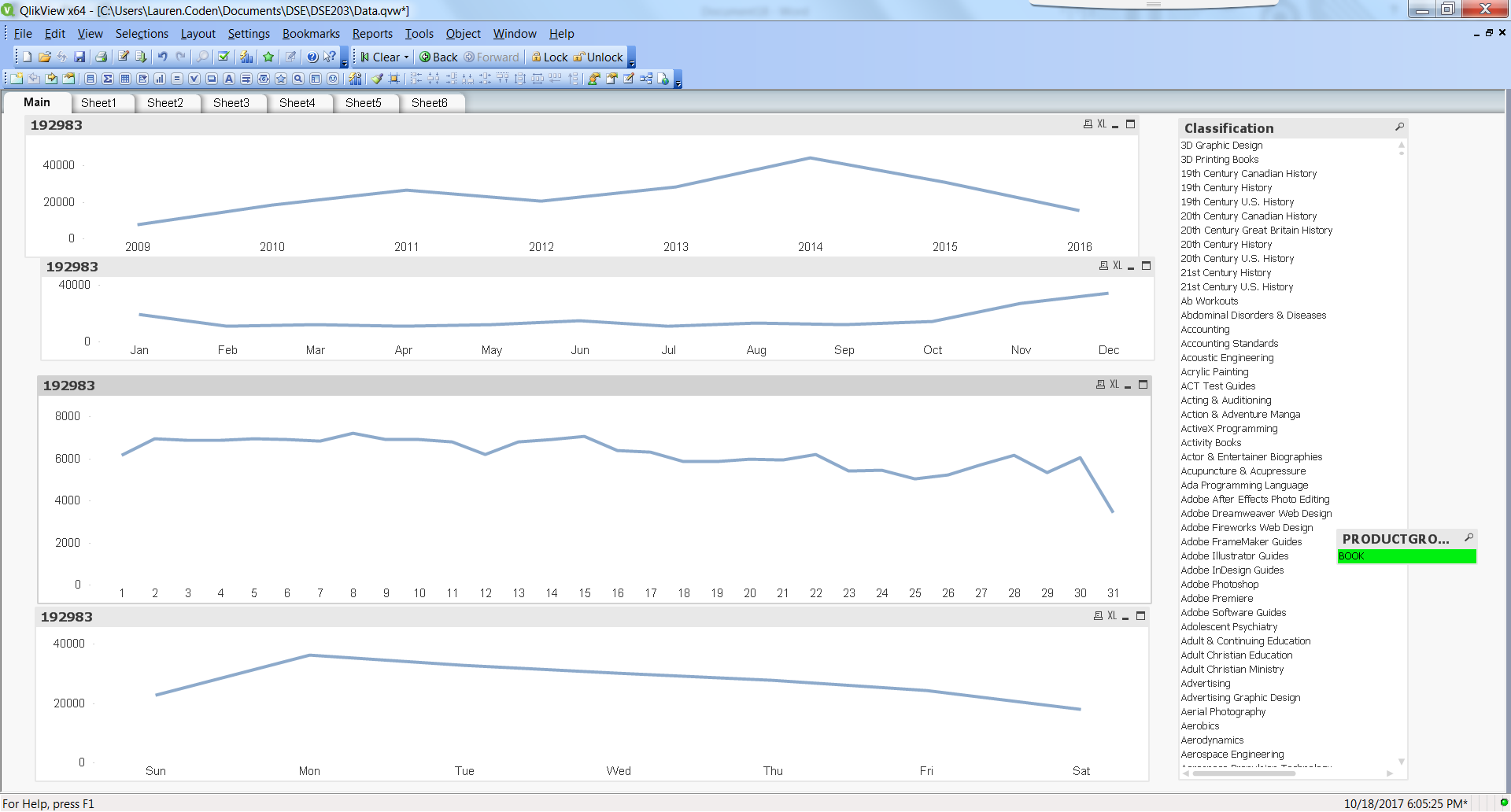
* Buckets per year? Year over year is consistent for the most part but peak in 2014. May want to model 2009 – 2013, 2014 and 2015 – 2016.



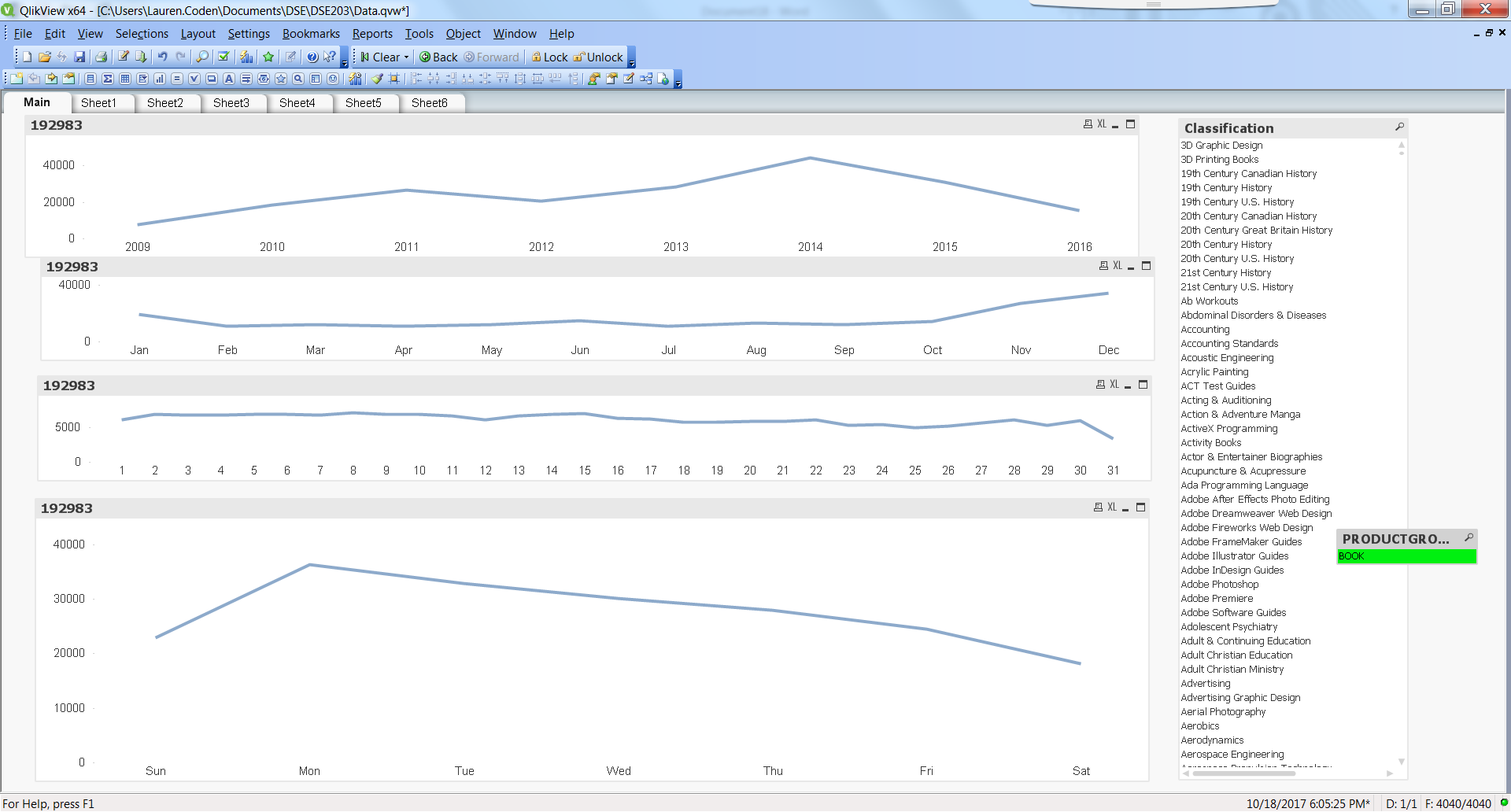
* Buckets per month? Sales around the holidays way up. May want to bucket holiday season (Nov, Dec, Jan) and non-holiday season (Feb – Oct)



* Buckets per day of month? Sales are slightly higher in the beginning of the month but mostly consistent. Count of orders at day 31 drops only because not all months have 31 days.

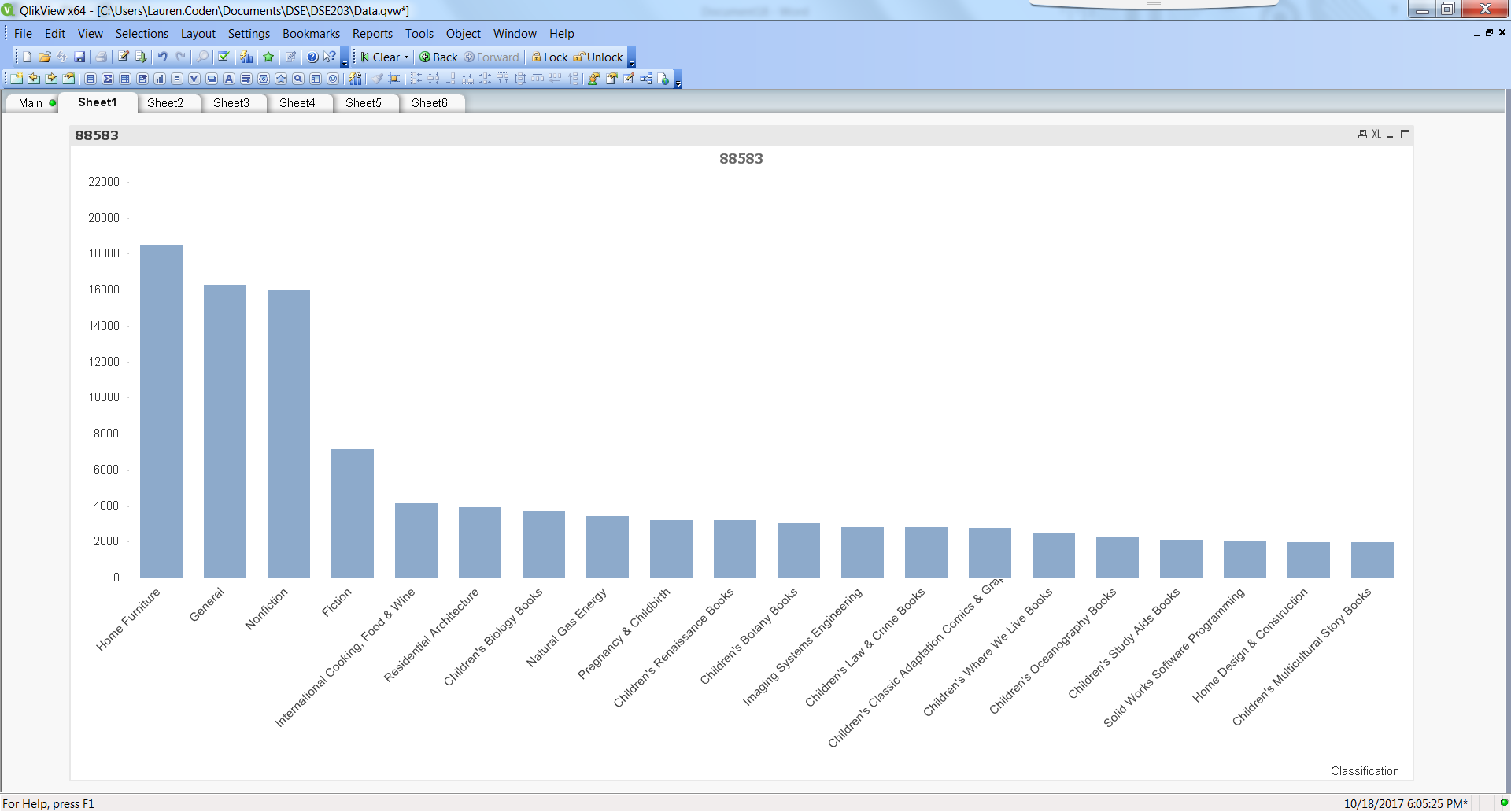


* Buckets per day of week? Mondays have the highest number of orders and the weekend lowest.

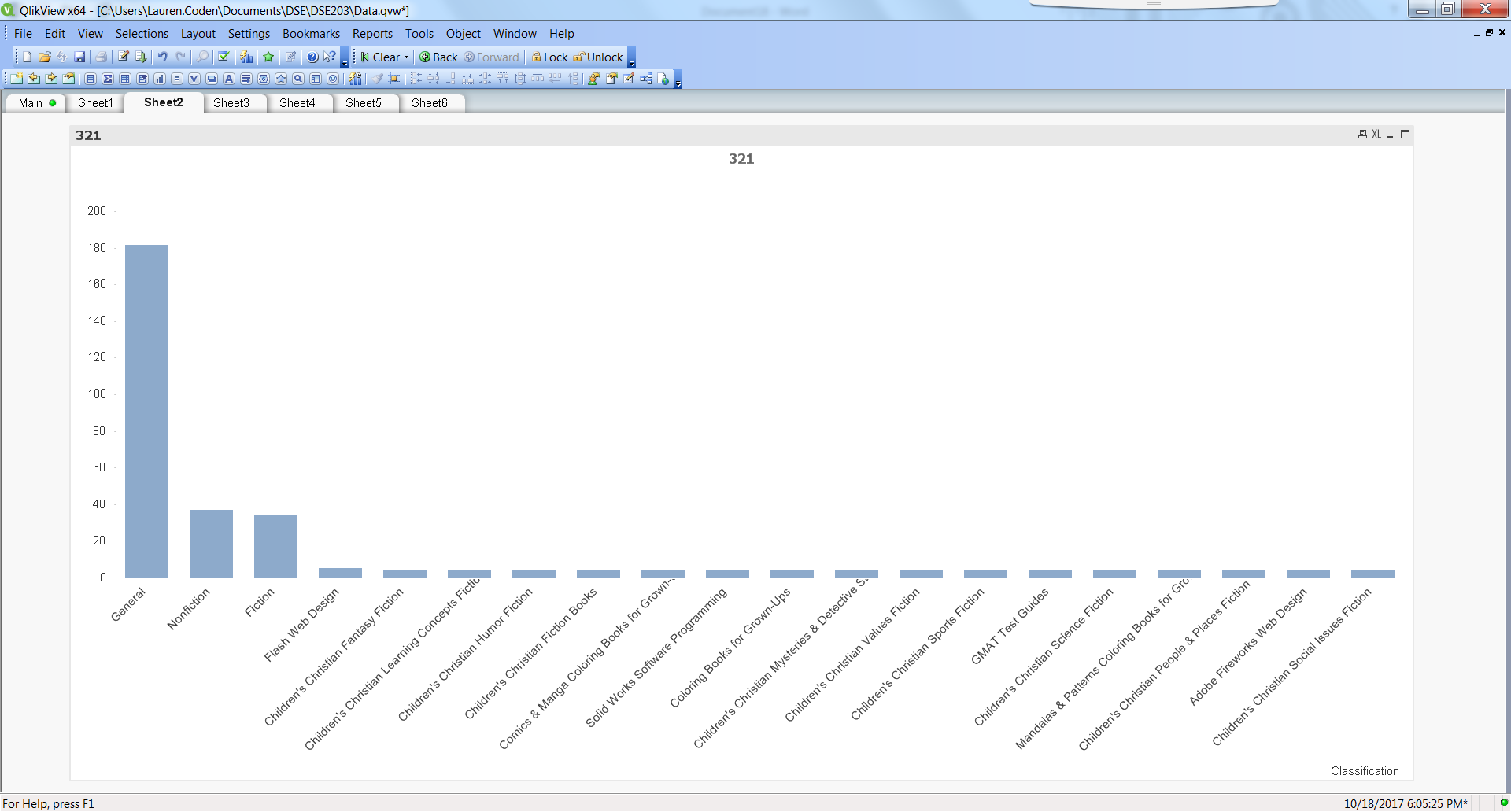


**Classification Buckets (Top 20):**

* Top 3 classifications are Home Furniture, General and Nonfiction. May be good to model these three in their own bucket.

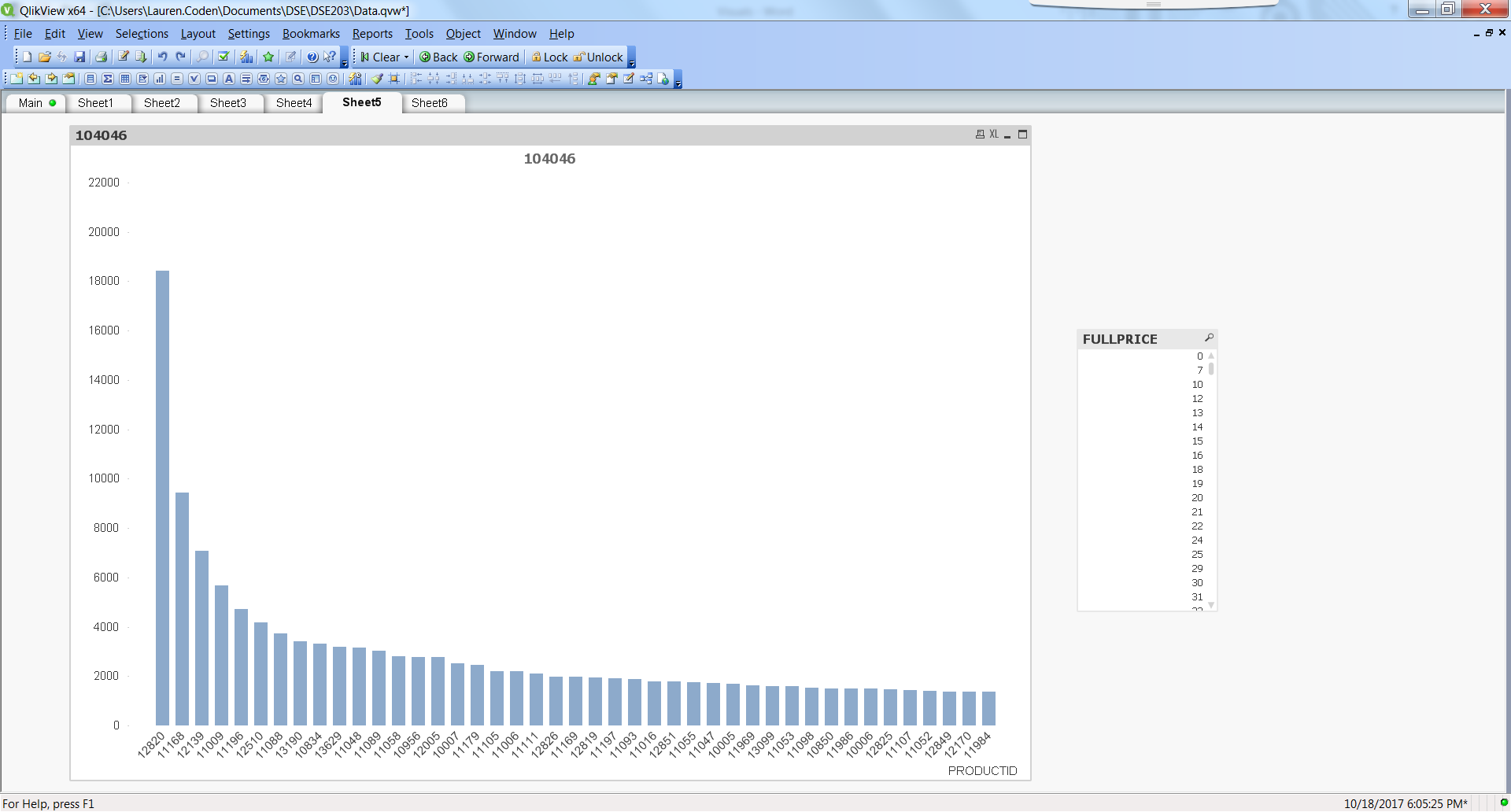


* The chart below shows a count of products per classification, which I made to verify whether the number of products in a classification is driving the above distribution of orders per classification. Home Furniture has high order count with low product count proving high performance.

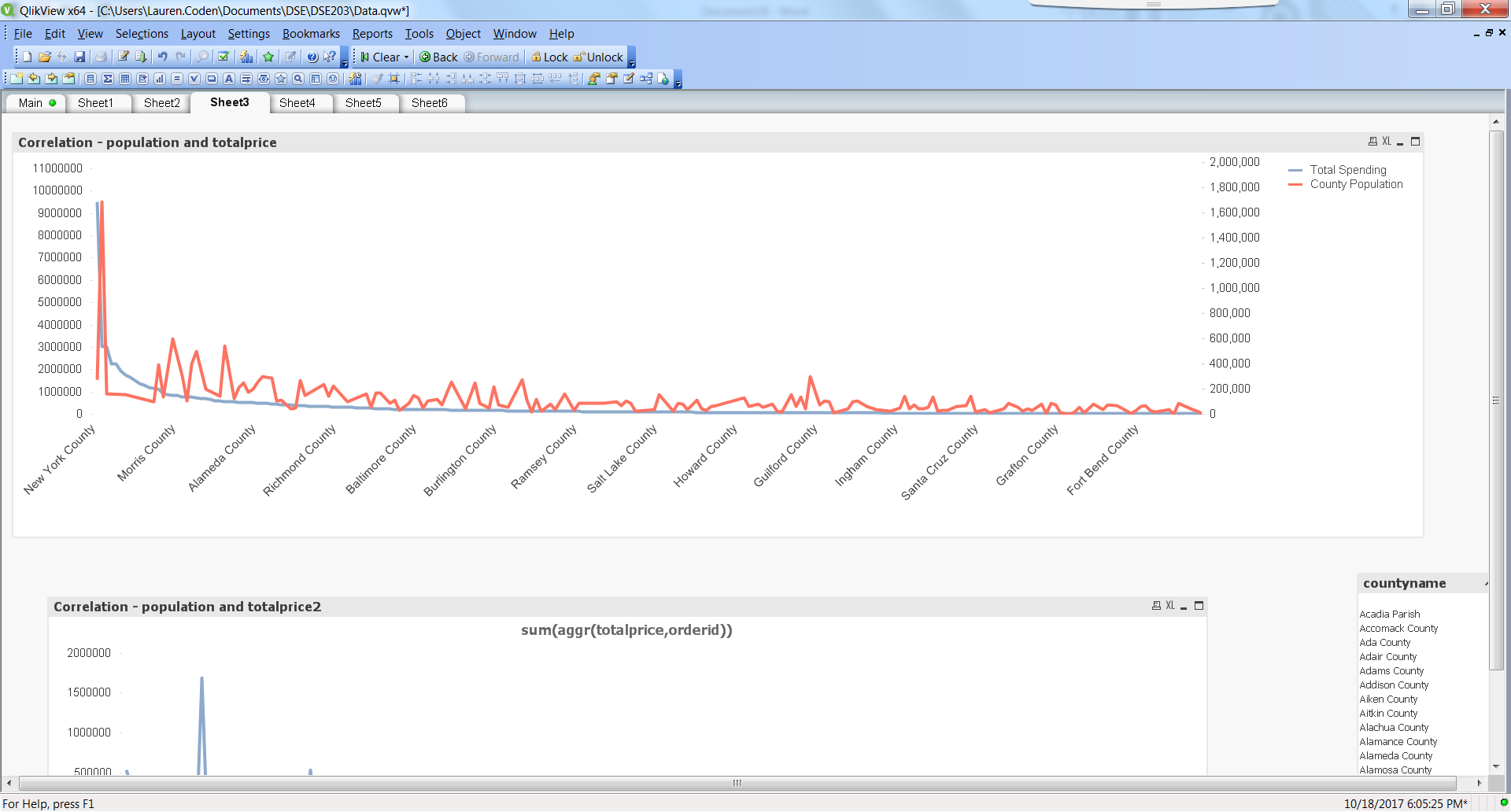


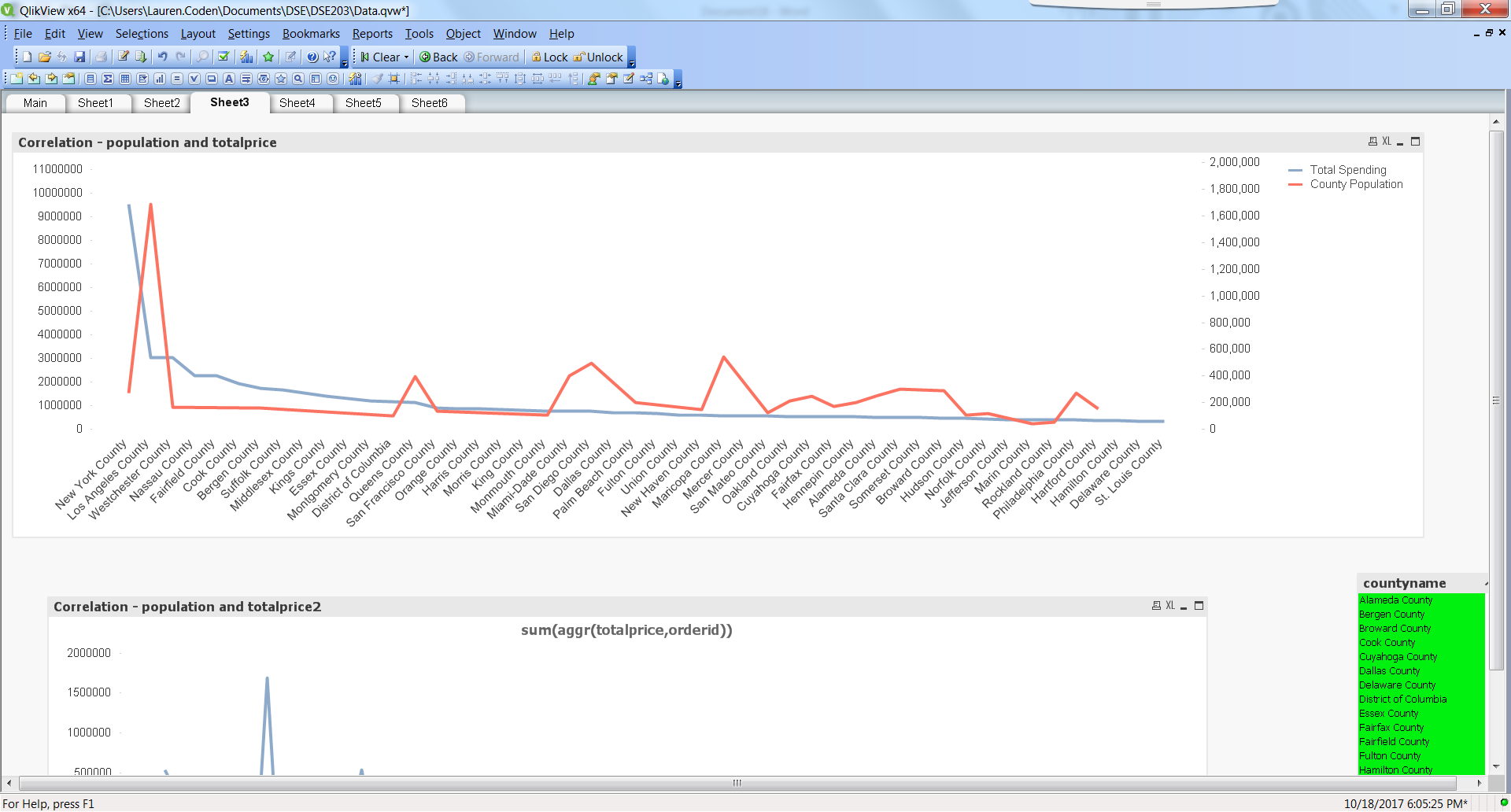
**Product Buckets:**

* Product 12820 has best sales and may want to be its own bucket.

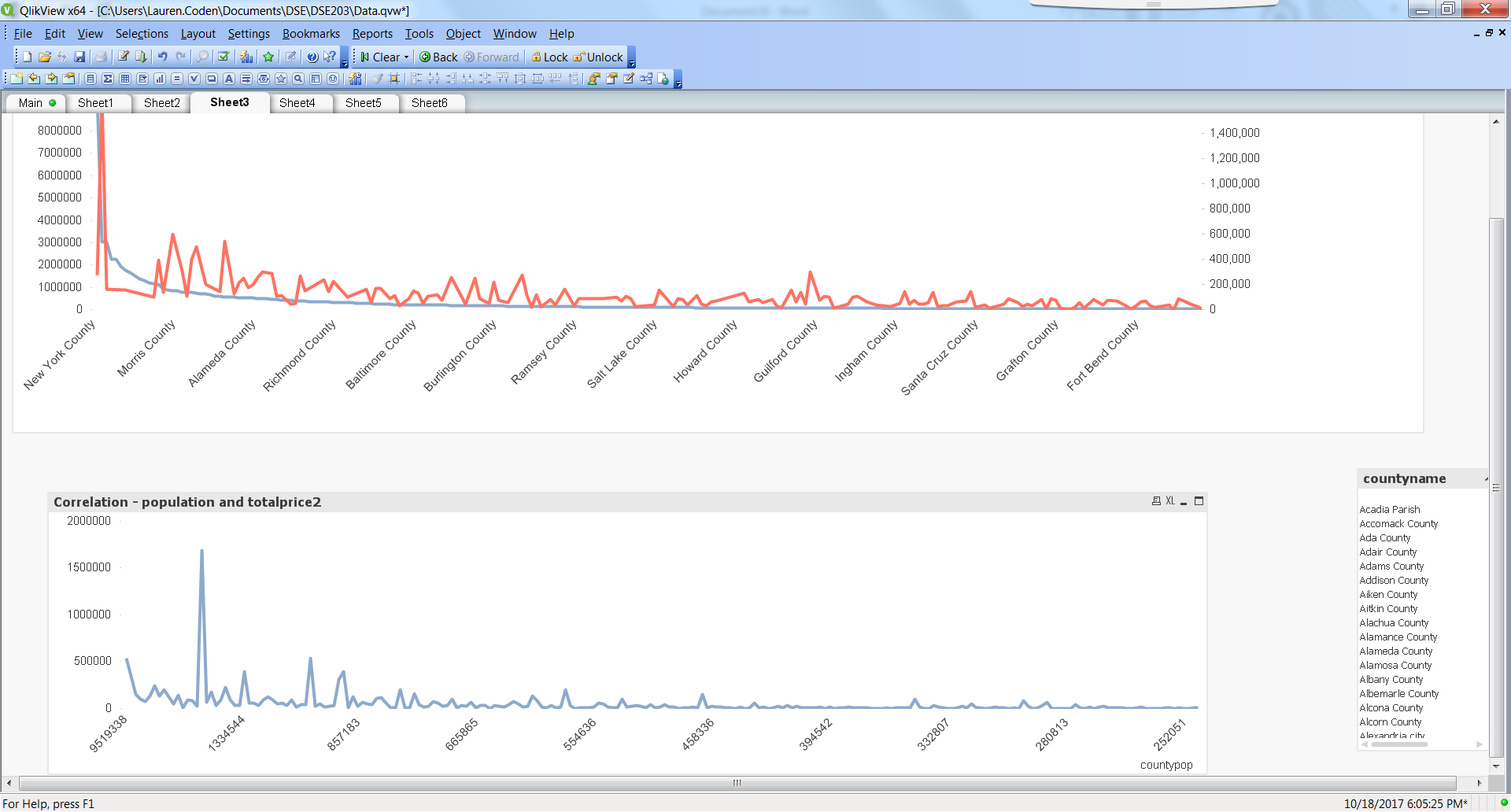


Correlation – population and total price

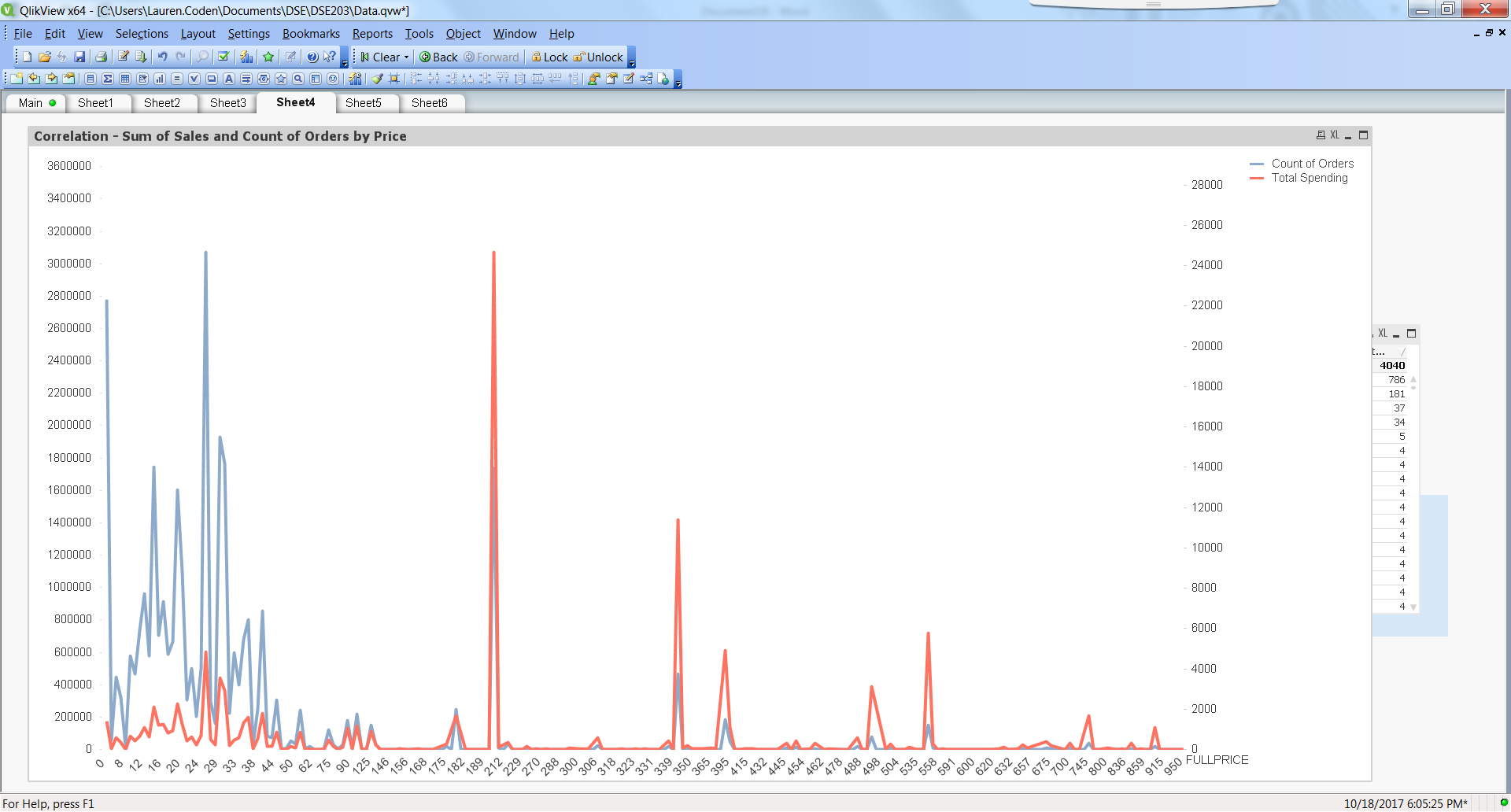




Below is Total spending by county population. First zip is Los Angeles – big population and somewhat big spending. Big spike is NYC – somewhat big population but HUGE spending. May want to model NYC separately.



Correlation – Sum of Sales and Count of Orders by Price



Gender Total Spending

* Blue is null. Males surprisingly have much higher sales.

