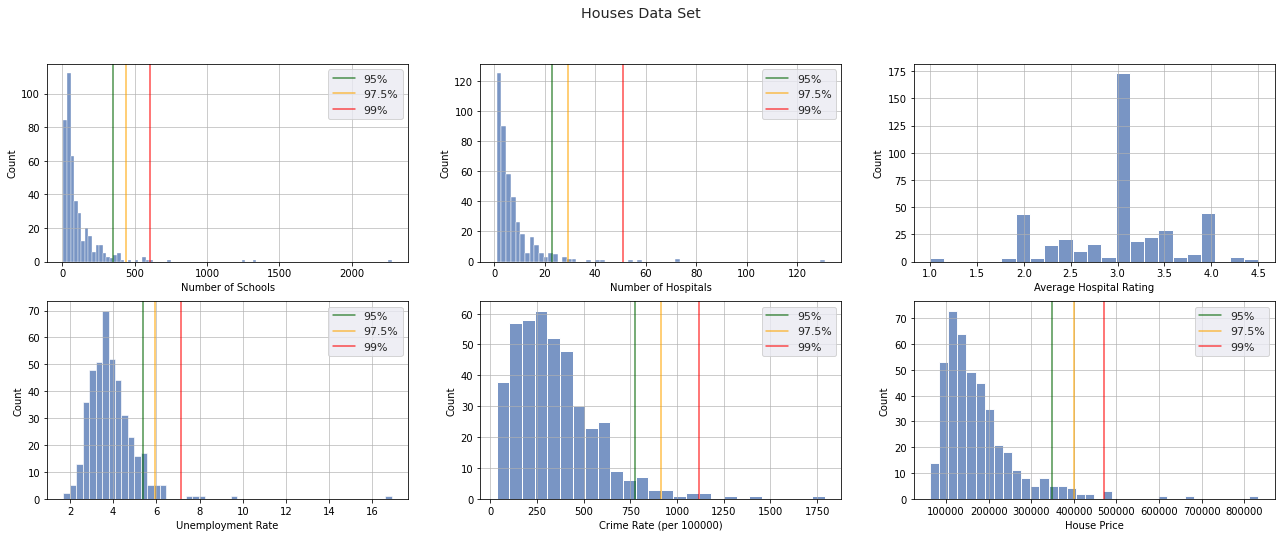
Modelación

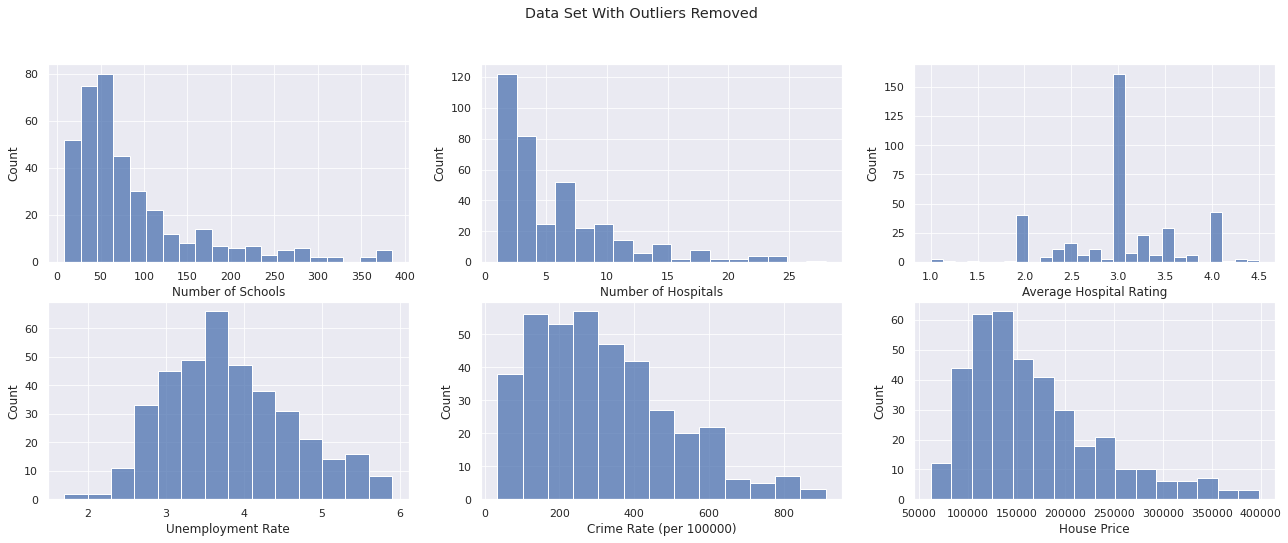
Como ya hemos mencionado en secciones anteriores, el objetivo particular de este proyecto es hallar una metodología que sea útil para valuar las casas dentro del mercado estadounidense. Lo que proponemos para completar este objetivo es ajustar un modelo cuya variable objetivo sea el precio de las casas y las variables candidatas a ser consideradas como predictores, es decir aquellas variables que formaran parte de los vectores de características para nuestras observaciones, sean las variables sociodemográficas que conforman la base de datos que logramos conseguir, limpiar e integrar en la sección anterior.

Una vez que tenemos lista nuestra base de datos, lo que siempre es recomendable hacer es un análisis exploratorio preliminar. En este análisis exploratorio preliminar, nuestro objetivo principal será seleccionar aquellas variables que contengan la mayor cantidad de información respecto al precio de las casas. La forma de determinar qué tanta información nos aportan estas variables candidatas depende del tipo de la variable, si es numérica y ordinal, o si es categórica.

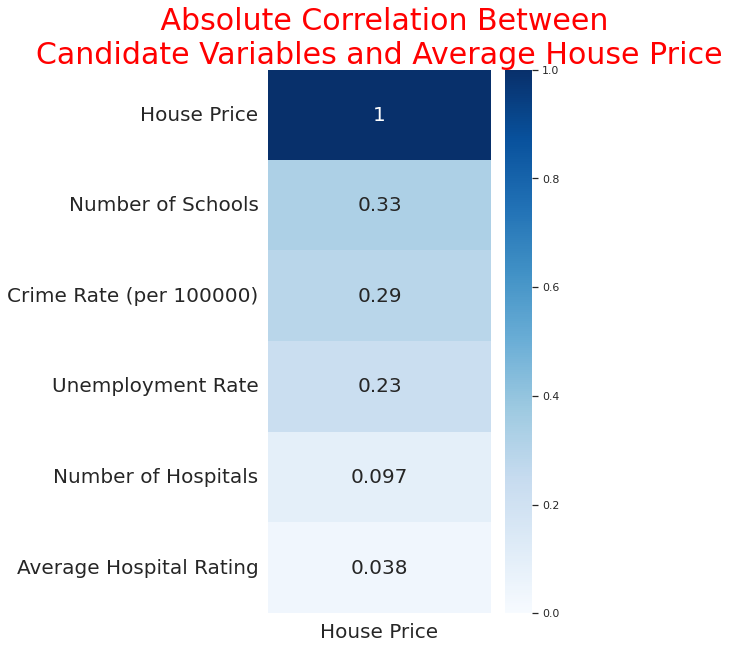
El primer paso en el análisis de las variables numéricas y ordinales suele ser el analizar los histogramas de cada una de éstas. Haciendo esto, es posible darse cuenta de 4 características, no necesariamente independientes una de otra, importantes correspondientes a las variables en cuestión: la distribución (si es normal, exponencial, uniforme, etc.), el rango (los valores que puede llegar a tomar la variable), los cuantiles (los cuales nos indican dónde se concentra cierta cantidad de los datos) y los valores atípicos (aquellas observaciones que presuntamente se alejan de un escenario normal). Por esta razón, decidimos hacer uso de esta poderosa herramienta, los resultados se pueden apreciar en la siguiente imagen:



Las líneas verticales de color verde, amarillo y rojo representan los cuantiles al **95%**, **97.5%** y **99%** respectivamente. Lo cual se puede interpretar como que el **95%** de los ejemplos dentro de nuestro data set se encuentran concentrados hasta la línea verde, el **97.5%** antes de la línea amarilla y el **99%** de los ejemplos se encuentran del lado izquierdo de la línea roja. Otra observación importante que debería hacerse de las gráficas es que muchas veces las líneas se encuentran en la mitad izquierda de las gráficas, esto quiere decir que existen valores que podrían considerarse atípicos. No es recomendable considerar valores atípicos al momento de modelar o sacar conclusiones respecto de las relaciones que guardan las diferentes variables, pues éstos únicamente representan ruido (representan información falsa o basura). Lo que decidimos hacer es considerar exclusivamente aquellos ejemplos que se encuentren del lado izquierdo de la línea amarilla para cada una de las variables. El resultado es el siguiente:



Con las distribuciones más estables, podemos continuar con nuestro análisis. El siguiente paso consta de encontrar las relaciones que existen entre las variables numéricas y ordinales candidatas y nuestra variable objetivo. El método clásico para estudiar estas relaciones es mediante la correlación y ésta puede observarse claramente con la ayuda de la siguiente visualización:



Lo ideal es que la correlación sea lo más cercana a la unidad. Mientras más cercano a 1 se encuentre el valor de la correlación implica que más información aporta la variable correspondiente respecto de nuestra variable objetivo. Por esta razón utilizaremos aquellas características que más correlación tengan con el precio de las casas. Consideraremos las variables: del **número de escuelas** dentro del condado, **el ratio de crímenes**, el **ratio de desempleo** del condado y el **número de hospitales.**

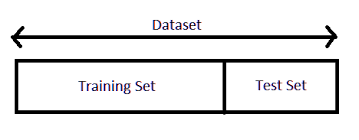
Ahora, para analizar la única variable categórica que tenemos (el estado en donde se encuentra la casa), nos apoyaremos de una prueba estadística que recibe el nombre de prueba de Kruskal. El objetivo de la prueba de Kruskal es determinar si una variable categórica aporta una cantidad significativa de información respecto de una variable numérica y ordinal. La hipótesis nula de esta prueba es que la variable categórica NO aporta información relevante. El resultado obtenido al aplicar dicha prueba para la variable correspondiente al estado en donde se encuentra la casa se encuentra en la tabla que se encuentra a continuación:

| Estadístico | P-valor |
| --- | --- |
| 172.25 | 8.8478 e-17 |

El valor más importante es el P-valor, el cual nos indica que, como dicho valor es menor a 0.05, podemos rechazar la hipótesis nula de la prueba con un 95% de confianza. Con esto concluimos que la variable correspondiente al **estado** donde se ubica la casa es una variable importante si se desea predecir el precio de la misma. Lo único que falta es encontrar una buena representación de dicha variable. En nuestro caso haremos uso del código FIPS, el cual en sus primeros dos digitos tiene la información del estado al que pertenece el condado, y además a condados cercanos se le asocia códigos cercanos, así habrá una relación de cercanía entre condados, esto nos puede ayudar ya que si un condado tiene valores de casas muy altos entonces esperaríamos que los condados vecinos tengan valores cercanos.

Ahora que hemos seleccionado las variables que utilizaremos para el modelado, conviene hablar sobre los modelos que utilizaremos. Como nuestro problema es uno de regresión, lo consecuente es usar algoritmos de aprendizaje de regresión. Los modelos candidatos que seleccionamos para intentar ajustarlos, probarlos, evaluarlos y compararlos son: regresión lineal, árbol de decisión y XGBoost. Respecto de las métricas que utilizaremos para evaluar y comparar los métodos anteriores son la y el MSE, esto debido a la naturaleza del problema y de los modelos que ajustaremos.

Para realizar todo este procedimiento de ajuste, prueba, evaluación y comparación de los modelos. Lo primero que debemos hacer es dividir nuestro conjunto de datos en dos subconjuntos: uno de entrenamiento y otro de prueba. El conjunto de entrenamiento constará del 80% del conjunto total mientras que el conjunto de prueba representará el sobrante 20%. La manera gráfica de verlo es la siguiente:



Posteriormente eliminamos el ruido de nuestro conjunto de entrenamiento, es decir, consideraremos únicamente las observaciones cuyas variables se encuentren antes de todas las líneas amarillas de la primera figura presentada en esta sección. Luego ajustamos los modelos sobre este conjunto preprocesado y, con estos modelos ajustados, estimamos los valores para el conjunto de ejemplos de prueba. Con estas estimaciones podemos calcular las métricas mencionadas algunos párrafos atrás.

**Resultados**

Los resultados se pueden resumir en la siguiente tabla:

| Modelo |  | MSE |
| --- | --- | --- |
| Regresión Lineal | 0.34 | 0.158 |
| Árbol de decisión | 0.18 | 0.180 |
| XGBoost | 0.554 | 0.102 |