Resultados

Parte I

1)

El primer paso es **recopilar los datos**, los cuales son el insumo principal que alimenta los modelos. En la practica los datos pueden ser de fuentes internas o externas, y de diversos tipos como series económicas, personas, transacciones, imágenes, vídeos, sonidos, mails, etc. Generalmente se representan en forma de matrices o tensores, donde cada columna es un numero que representa una medida o una categoría ordinal.

El segundo paso es la **exploración y preparación de los datos**, este paso tiene dos elementos en el primero la idea es tener la máxima intuición sobre el posible modelo, para eso exploramos los datos extrayendo la mayor cantidad de información posible que permita establecer una guía hacía la elección del modelo correcto. El segundo es la preparación de los datos, para esto nos basamos en la información obtenida a través del análisis exploratorio y lo que dice el estado de la investigación actual sobre el fenómeno a modelar.

El tercer paso es **entrenar el modelo**, aquí la clave es la elección correcta del modelo, ya sea si es un problema de aprendizaje supervisado de regresión o clasificación, o uno no supervisado. El análisis exploratorio y la investigación actual nos pueden dar una referencia hacia cual es un buena primera aproximación. Es importante dado que asumimos que este sera el proceso que mejor representa a los datos que buscamos modelar.

El cuarto paso es la **evaluación del modelo**, aquí nos interesa saber que tan bien lo hizo el modelo, en relación con la experiencia. Esto implica desarrollar medidas que nos permitan elegir entre dos, o mas, tipos de modelos. La definición de que es o no un buen resultado vendrá dado por las necesidades del negocio o de la investigación y las características de los datos.

Finalmente, el quinto paso es **mejorar el modelo**. La mejora del modelo implica ajustar el modelo con nuevas variables, agregar datos adicionales, profundizar el análisis exploratorio. Básicamente si no estoy conforme con mi modelo me debo preguntar en cuales de los pasos podre estar fallando y evaluar cual es elemento que nos falta incorporar.

2)

Aprendizaje Supervisado: Es una técnica que nos permite obtener una función para modelar una variable deseada a partir de un conjunto de datos de entrenamiento. Pueden ser regresiones, lo que implica que la variable modelada es continua. O ser de clasificación, donde la variable modelada es discreta.

Algoritmo: Regresion LASSO.

Paper : Industry Return Predictability: A Machine Learning Approach, Rapach 2018.

Resumen : Ocupa variables rezagadas de activos y a nivel industrial para predecir los retornos de la industria.

Data Paper: Link a los Datos

Apredizaje no Supervisado: Se busca modelar una función que permitan obtener una idea de la relación existente en los datos, por lo tanto nos estamos a priori interesados en modelar solo una variable si no conocer mejor la relación estructural de los datos.

Algoritmo: Hierachical Clustering.

Paper: Building Diversified Portfolios that Outperform Out-of-Sample, Lopez de Prado 2019.

Resumen: Ocupa los cluster para agrupar los retornos de los activos, y así obtener una matriz de covarianza robusta para predicciones fuera de muestra.

Ejemplo: Link al medium

Data Ejemplo: Link a los Datos

Parte II

Introducción

Todos el código que incluye gráficos, tablas y los modelos se encuentra **aquí**. Se incluirán solos los principales resultados.

Son 41188 observaciones, donde 11 corresponde a variables categóricas y 9 son numéricas. En promedio la edad de los clientes corresponde a 40 años, al menos el 50% de los datos se encuentre entre 32 a 47 años dado los cuantiles. La duración de las llamadas parece concentrarse en el lado izquierdo de la distribución y existe la presencia de outliers, la media es de 258 seg. y y tiene un máximo de 4918 seg. mientras el 75% de los datos es menor a 319 seg.. En campaña y previo ocurre algo similar con la mayoría de clientes llamados pocas veces y algunos outliers. Los indices económicos parecen no tener outliers. Al igual que el numero de empleados. La mayoría trabaja como Admin, después viene los blue-collar seguidos por los technician en tercer lugar. 24928 se encuentran casados y 11568 solteros. Una porcentaje alto tiene educación universitaria o high school. Solo 3 han caído en default, 6248 tienen prestamos. Para contacto es mayor las personas que escogen celular, y en general el ultimo contacto se hizo en mayo. El día tiene una distribución cercana a una uniforme, pero existe una leve concentración los días lunes y martes. La mayoría es primera vez que ha sido contactado, y solo 1373 han contratado el producto después de la campaña. Finalmente, cercano al 89% de los clientes no tiene el producto. Dentro de todas categorías posibles solo en dos es mayor los que tiene a los que no tiene el producto, y esto es en la categorías asociada a las campañas del mes de marzo y a las que ya tuvieron éxito en la llamada anterior. En general no podemos apreciar diferencias a simple vista en cuanto a diferencias entre la contratación del producto y el nivel de las variables numéricas. Una excepción es la variable campaña, dado que es notorio que para llamadas mas de 15 o 20 veces los clientes tienden a ya no contratar el producto.

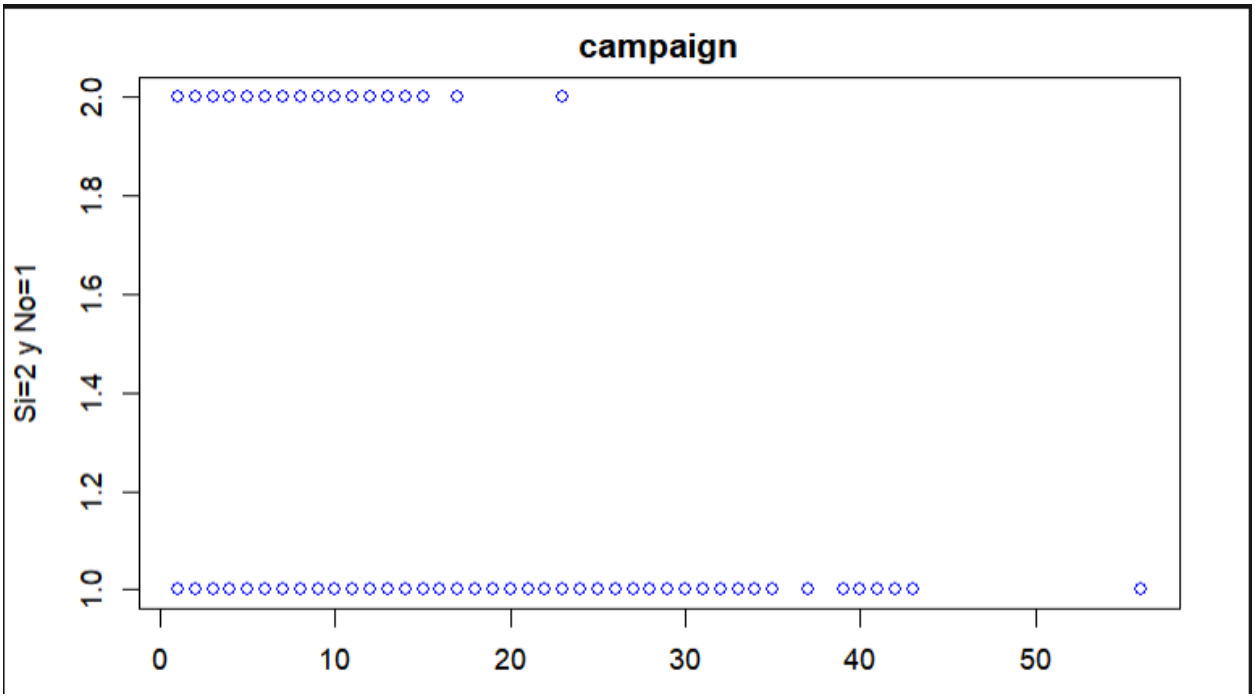


Figura 1: Cantidad de llamados, segun si contesta o no.

	no	yes
failure	3647	605
nonexistent	32422	3141
success	479	894

	no	yes
apr	2093	539
aug	5523	655
dec	93	89
jul	6525	649
jun	4759	559
mar	270	276
may	12883	886
nov	3685	416
oct	403	315
sep	314	256

Figura 2: Success y March son las unicas dos categorías dentro de toda la data, donde el si es mayor al no.

Modelos Iniciales

KNN Se escogió el algoritmo KNN y el decision tree(DT) para modelar si es que los clientes obtendrán el producto. Para ambos algoritmos los datos fueron normalizados y se ocupo validación cruzada con k=30 como método de remuestreo. Se escogió un conjunto de entrenamiento del 80% de los datos dejando el 20% para el testeo.

La Confusion Matrix y precision para el algoritmo KNN son las siguientes:

knn_model_result	no	yes
no	7065	549
yes	245	379

[1] "Accuracy:" "0.904"

Obtuvimos una mejora del 1% si lo comparamos con haber supuesto ninguno tenia el producto. Hay mas errores cuando el modelo predice si y es no que en el caso contrario. Los datos predicho se encuentran bien balanceados.

KNN: Mejora Como **mejora** propusimos un repeated-CV donde hicimos 5 repeticiones, pero el modelo arrojo un resultado similar en la cantidad de vecinos por lo que no hubo mejora en la performance fuera de muestra.

Decision Tree La Confusion Matrix y precisión para el algoritmo DT son las siguientes:

DT_result	no	yes
no	7047	451
yes	263	477

[1] "Accuracy:" "0.913"

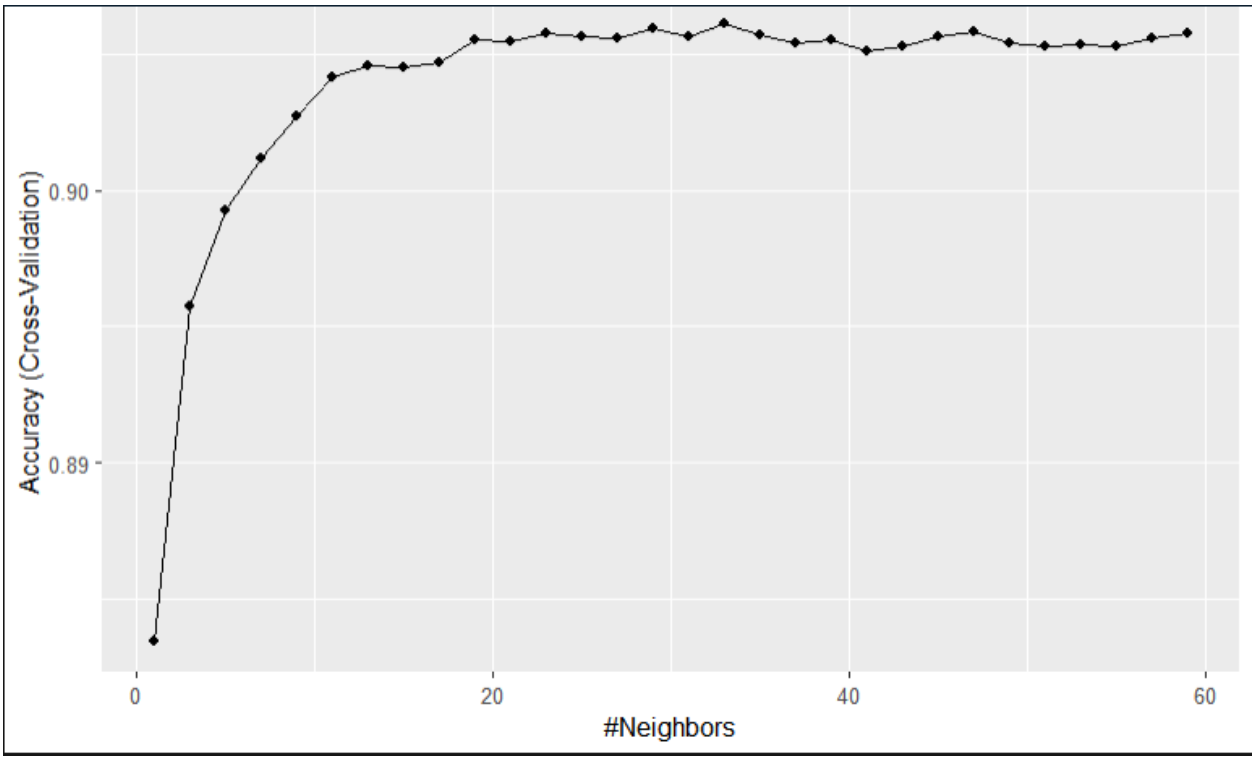


Figura 3: CV k=10, KNN

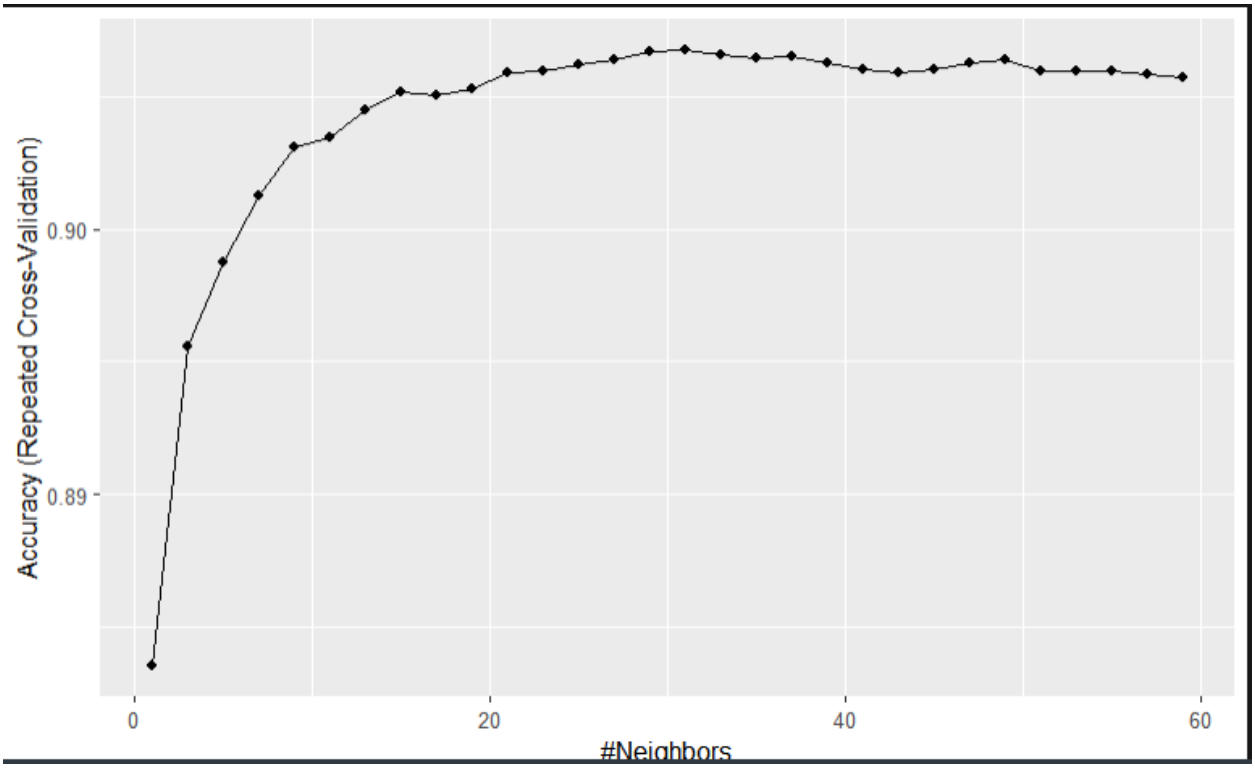
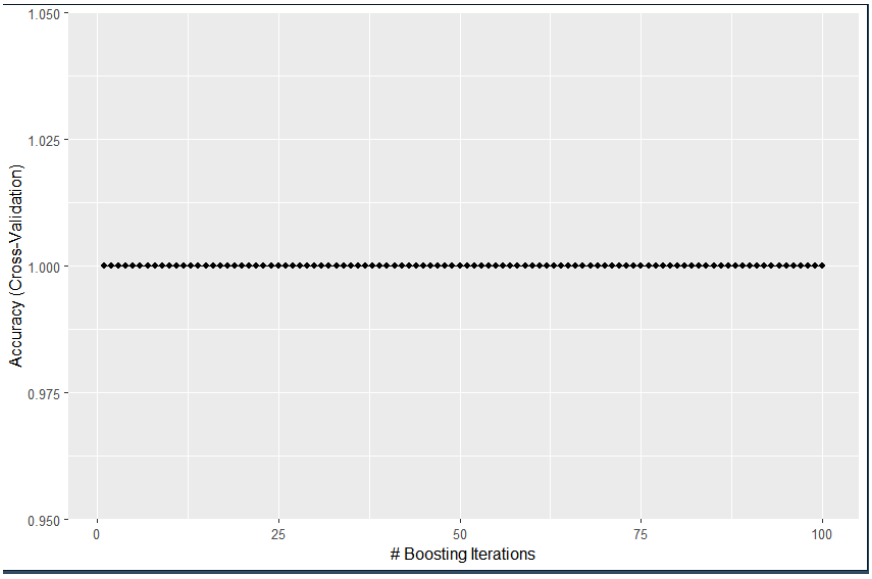


Figura 4: RCV k=10 y r =5, KNN

Se escogieron 100 iteraciones, dado que el CV nos mostraba una precisión perfecta en el set de entrenamiento independiente de la cantidad de iteraciones que usáramos.



Obtuvimos una mejora del 2% si lo comparamos con haber supuesto ninguno tenia el producto. Nuevamente ocurre la concentración de errores cuando es no pero el modelo predice si. la principal mejora es cuando en el aumento de los si.

DT: Mejora La mejora consiste en usar Random Forest compuesto de 1000 arboles que debiese permitirnos obtener un mejor resultado.

```
    y_pred
       no  yes
  no  7026  284
  yes  425  503
[1] "Accuracy:" "0.914"
```

Obtenemos una mejora marginal, si bien la precisión de la categoría no cae aumento la de la categoría si, aumentando en el agregado.

Parte III

Todos el código que incluye gráficos, tablas y los modelos se encuentra **aquí**. Se incluirán solos los principales resultados.

Descripción de variables.

Numéricas

- MonthSold: Mes en cual fue vendida. Va del 1 a 12, donde cada numero es un mes y tiene una media y mediana de 6 meses. Podría existir un patron de estacionalidad que haga que las ventas se mayores en algunos meses

- Size(sqf): Tamaño de la vivienda. Va desde los 135 a 2337sqf con una media de 955sqf. A mas tamaño debiese ser mas cara.
- Floor: La cantidad de pisos. Van desde 1 hasta 43, por lo tanto existen departamento. La mediana es de 11 departamentos. En general una casa con mas pisos es mas cara, pero en el departamento el efecto agregado no debiese impactar en el precio de uno. O puede ser negativo, mas departamento lo percibo como algo malo.
- N_Parkinglot(Ground): Cantidad de estacionamiento. Desde 0 a 713 y con media de 195. Mas estacionamiento, en general, implican mas caro.
- N_Parkinglot(Basement): Cantidad de estacionamiento. Desde 0 a 1321 y con media de 570. Mismo argumento punto anterior.
- N_manager: Cantidad de administradores de la vivienda. Una media y mediana de 6 y se ubica entre 1 a 14. No debiese haber una relación tan directa en el precio.
- N_elevators: Cantidad de ascensores. Va de 0 a 27 con una media y mediana de 11.
- PublicOffice: entre 0 y 7 oficinas publicas, con una media de 4 y mediana de 5. Mas oficinas publicas pueden indicar zonas con mas comercio pero menos privacidad.
- Hospital: entre 0 y 2. Creo que aquí la diferencia esta en 0 o mas, no veo que debería existir un diferencia significativa entre tener 1 o 2 hospitales. Pero si entre 0 y 1.
- DpartmentStore y Mall: Entre 0 y 2, con mediana de 1.
- Otras: Otras facilities cercanas entre 0 y 5.
- Park: Entre 0 y 2. Esperaríamos que los parque aumentaran el valor de una propiedad.
- Elementary: Escuela elementaria, entre 0 y 6 con media y mediana de 3. Una vivienda sin escuelas cercanas en general debiese ser malo. Por otro lado, el efecto debiese ser decreciente dado que tener entre 4 a 6 debiese dar un poco lo mismo.
- En general lo mismo es valido para los otro niveles de escuela como Middle,High y la universidad.
- Finalmente están los totales, primero de las facilities donde esperaríamos que a mas sea mejor la vivienda, y los mismo para el total de escuelas.

Categoricas

- HallwaType: Existen 3 tipos corridor,mixed y terraced. Este ultimo es que contiene mas vivienda con mas del doble que mixed, que ya es mas de doble que corridor.
- HeatingType: Solo 300 tienen central heating, la mayoría tiene individual heating.
- ApotManageType: La mayoría es administrada por un fideicomiso y solo 349 son administrados por cuenta propia.
- TimetoBussStop: La mayoría se encuentra en el rango de 0- 5 min, menos de un cuarto se ubica entre 5-15 min, en particular en el tramo 10-15 min solo se ubican 55 personas.
- TimetoSybway: Los que tienen se ubican entre 0-20 min, existen 238 viviendas donde no existe metro cerca, la mayoria se ubica entre los 0-10 min.
- SubwayStation: La estación mas cercana a las viviendas es Kyugbuk uni hospital, seguida por Myung-duk. Seguidas en orden por Banwoldang, Bangoge, Sin-nam, no_subway_nearby y Other.

Dentro de las posibles relaciones que notamos en los gráficos bivariados notamos que los departamentos mayores de 40 pisos tienen 2 departament store cerca, y los de treinta al menos 1. Para los otras cantidades pueden variar,Similar lo que ocurre el los parques. En hospitales y mall ocurre algo similar el 1 es el mas común y departamentos mayores de 30 la probabilidad de que tenga 1 es muy alta. Pero que tengan 0 o 2 es muy baja. Pisos mayores a 35 pisos requieren 14 managers. En Parking pasa que en Basement se concentra en la parte de la derecha de la distribución el numero de estacionamientos, y en el Ground se concentra a la izquierda. Existe una evidente relación positiva entre y el numero de manager y n_APT. Siendo las relaciones mas interesantes que se desprenden del análisis bivariado.

Método del codo.

Primero normalizamos los datos. Posterior a esto corremos el metodo del codo para escoger el optimo de cluster, con un maximo de 5. Los tres métodos muestra cosas distinta. Nos quedaremos con el tres que en el general parece hacerlo mejor.

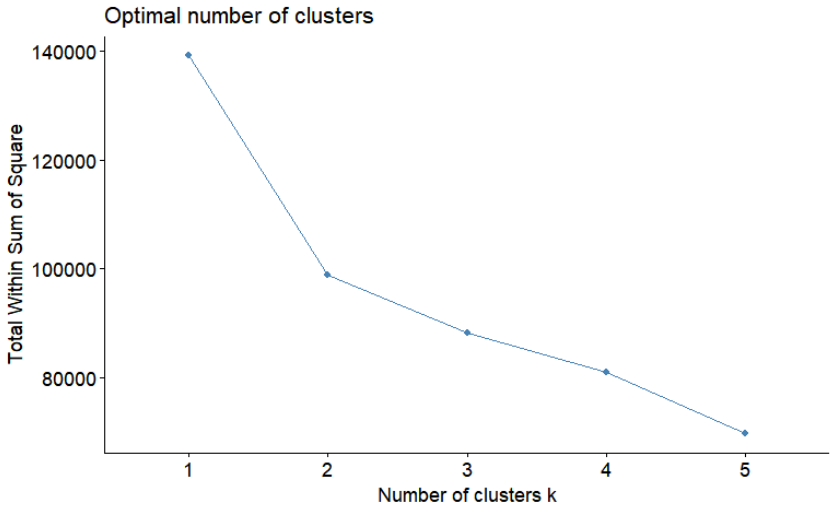


Figura 5: WSS

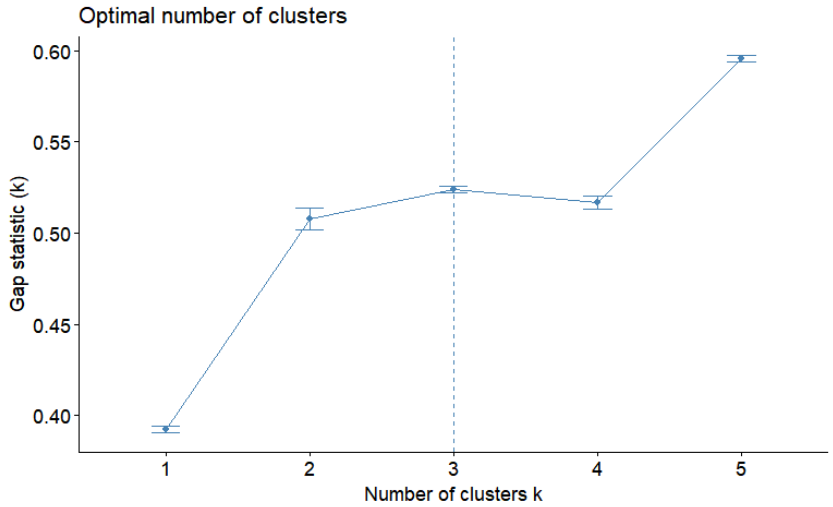
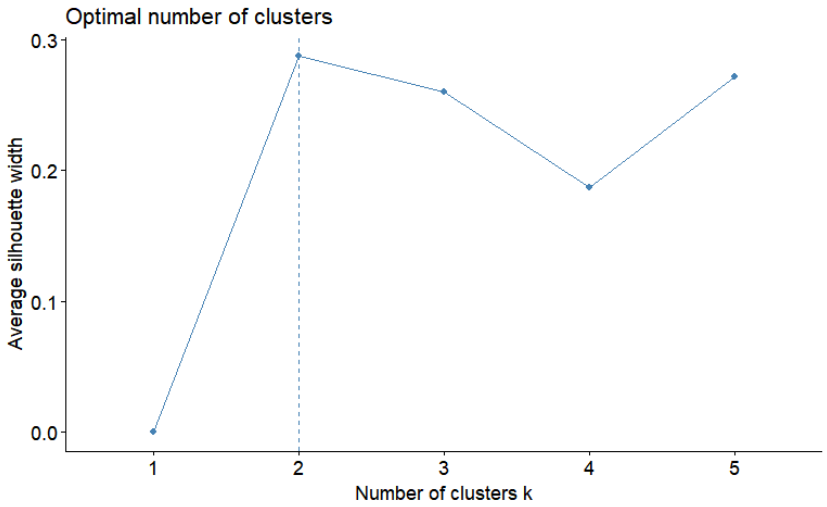


Figura 6: Gap



Se obtuvo como resultado tres cluster. De tamaño 1189, 2783 y 1919.

Si revisamos las medias de cada cluster tenemos lo siguiente. El Monthsold afecta mayormente al cluster 1, el tamaño al cluster 3, el piso a la cantidad de pisos, los estacionamiento en el Ground y Basement afectan principalmente al 1 pero con diferente signo. El APT tiene una relacion parecida con el cluster 1 y 3 pero contraria con el 2. Similar a lo que ocurre con los manager pero es evidente que afecta mas al cluster 1. La cantidad de ascensores afecta principalmente al cluster 1 y de forma contraria al 2 pero con menor magnitud. Oficina publicas afecta al cluster 3 al igual que los hospitales y el ETC, el efecto sobre el cluster 1 y 2 es similar. Departament Store afecta principalmente al 1 y 3 con distinta direccion. Las escuela afecta de forma similar con el cluster tres con media positiva y los otros dos con negativa. Los totales afectan de forma similar positivo al cluster 3 y negativo al cluster 1 y 2.

Ahora se realizara el análisis gráfico para profundizar el las relaciones.

Al revisar los boxplot notamos que el tamaño de las viviendas es mas grande en los cluster 1 y 2 y menor en el 3. La cantidad de pisos es mayor en el 1 que en el dos y tres. En las variables categorías los 300 con central heating se concentra en el cluster 2. Al igual que los que administran ellos mismo la propiedad. Otra categoría que responde a algo similar es que el cluster 2 esta entre 10 - 15min. Notamos que el grupo 1 se encuentra concentrado entrono a la estación Kyungbuk_uni_hospital. Lo que nos permite concluir que en el cluster uno corresponde a viviendas de mayor precio, el cluster dos de precio intermedio y el cluster 1 de precios bajo.

Si miramos el boxplot incluyendo a las ventas notaremos que efectivamente esto es así.

Conclusión

Notamos que el aprendizaje no supervisado nos permitió encontrar cluster, que si bien podíamos haberlos señalado de forma explicita al incluir las variables ventas, que fueron encontrados por medio de sus características. Lo que habla del potencial como herramientas para aprender estructura y relaciones en los datos que a priori no son conocidas a través de los métodos de aprendizaje no supervisado.

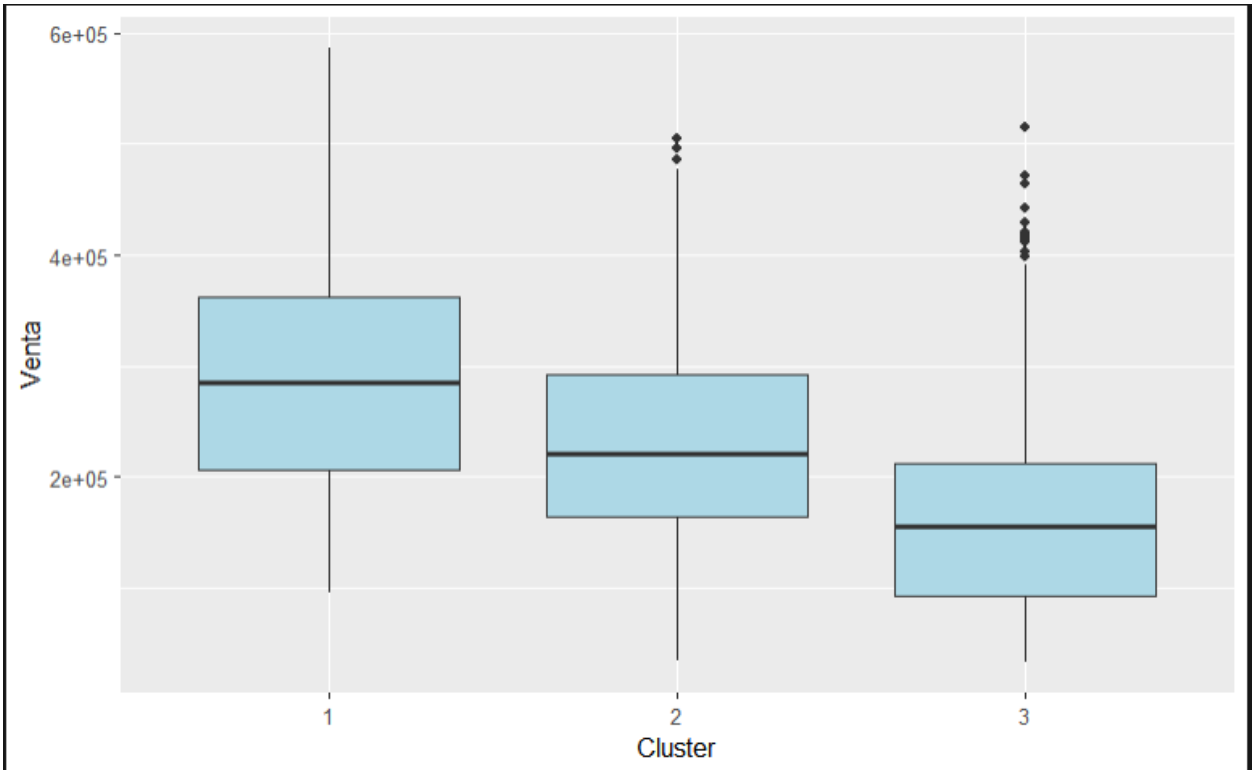


Figura 7: boxplot Vts vs Cluster

```
[1] "##########"      "HallwayType"     "###############"

             1    2    3
  corridor   0  332  305
  mixed      0  279 1411
  terraced 1189 2172  203
[1] "##########################"
[1] "##########"      "HeatingType"     "###############"

                      1    2    3
  central_heating     0  300    0
  individual_heating 1189 2483 1919
[1] "##########################"
[1] "##########"      "AptManageType"   "###############"

                        1    2    3
  management_in_trust 1189 2434 1919
  self_management        0  349    0
```

Figura 8: Table: Algunas Varibles Categoricas