UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Proyectos II, integración, calidad y análisis exploratorio de datos

Segunda presentación:

Integración de datos

Introducción

Antes de cada HITO de presentación, debemos rellenar estas fichas y presentarlas a través de PoliformaT, en la tarea que indiquen los profesores. Cada equipo de trabajo presenta las mismas fichas. Sólo será necesario que las suba uno de los componentes del equipo.

Nombres y apellidos de los autores:

Daniel	Garijo Abarca
Javier	Luque Sáiz
Pablo	Parrilla Cañadas
Jose	Valero Sanchis
Qilu Diana	Wu

2. Interés y alcance del proyecto.

Contesta a las preguntas que plantea este formulario.

2.1. Explica el objetivo principal de tu proyecto ¿qué presenta este estudio?

Antes de hablar del objetivo principal y centrarnos en el tema de nuestro grupo, podemos distinguir que nuestros objetivos formativos al hacer este trabajo son aprender a trabajar en equipo y adaptarse a las fechas límite que se imponen para realizar los hitos. Así pues, también buscamos instruirnos en la práctica de lenguajes informáticos como son R, Matlab o Python para poder realizar análisis precisos y variados, pero que sean claros para el receptor.

Por una parte, la finalidad principal de nuestro proyecto es poner en estudio la demanda de las bicicletas que se utilizan de las estaciones de Valenbisi para que, de esta forma, podamos realizar ciertos análisis y predicciones acerca de estas y emplear las conclusiones que obtengamos en nuestro día a día cuando las utilicemos para nuestro uso personal, o queramos obtener información acerca de donde debería reponer más seguido las bicicletas la empresa de reparto. Es decir, buscamos mostrar las capacidades que tenemos para realizar análisis acerca de las bicicletas que hay por Valencia.

Por otra parte, respondiendo a la pregunta, este estudio presenta una serie de pautas que estamos trabajando para llegar a las conclusiones mencionadas previamente, que son identificar las estaciones de bicis de esta marca alrededor de los barrios de Valencia, así como los lugares de interés que tengan estas a su alrededor para hacernos una idea de la demanda de bicicletas que se pueda ocasionar. Además de las pautas, este estudio presenta también dificultades/inconvenientes, como son el esclarecimiento de todas las variables de las bases de datos que disponemos a la hora de trabajarlas (por ejemplo, con la variable "updated_at" del dataset de Valenbisi), y la posibilidad de realizar numerosos análisis para poner a prueba nuestra capacidad de programar como es, por ejemplo, el clustering.

2.2. Explica para qué y para quién podría ser de utilidad este estudio

Por ejemplo, queremos observar y caracterizar la influencia de los medios de las redes sociales en la compra de ciertas mascarillas.

Puede interesar al gobierno porque..., a fabricantes porque…, a farmacias porque..., al público en general porque...

Como bien se menciona en el ejemplo, nuestro objetivo es observar la influencia que tiene la demanda de bicicletas de Valenbisi según el barrio de Valencia en el que se encuentre, así como las condiciones que pueden hacer que dicha demanda aumente o descienda (ya sea las precipitaciones que haya en esas fechas, la densidad de población que haya en ese barrio específico...).

La realización de este proyecto puede resultarle interesante, a la par que útil, a la empresa que se ocupe de distribuir y recargar las bicicletas y a los usuarios que utilicen diariamente el servicio Valenbisi, ya que les permitiría saber en qué zonas exactas se utilizan un mayor número de bicicletas y si puede haber más probabilidad de que tengan un borne libre o una bicicleta sin ocupar en una estación en concreto.

2.3. ¿Por qué piensas que es novedoso? ¿has visto estudios similares?

Pensamos que este estudio es novedoso debido a la innovación de los datos con los que trabajamos, que son actualizados cada quince minutos de forma diaria por parte de un docente de la UPV, de forma que cualquier estudio relacionado con este tema (la demanda de Valenbisi) estará anticuado en relación con los datos que se tratan en este proyecto. Así pues, no hemos visto estudios similares a este, pero sabemos que se habrán realizado estudios que no traten el tema de las bicicletas en específico, si no que esté relacionado con el ámbito del transporte (como por ejemplo el autobús, el metro...). Además, cabe destacar que es posible que se hayan realizado estudios sobre bicicletas, pero no sobre las de servicio de Valencia, si no los servicios en otras ciudades como pueden ser Madrid (*Bicimad*) o Barcelona (*Bicing*).

2.4. Alcance (objetivos definitivos del proyecto)

Define los objetivos del análisis de datos de tu proyecto. Se deben presentar 5 objetivos de análisis.

Con respecto al ejemplo anterior, posibles objetivos de análisis podrían ser:

- Observar la tendencia de ventas de mascarilla por marca desde el inicio de la pandemia hasta la fecha de…
- Estudiar la relación entre las marcas de mascarillas y los datos de…
- Estudiar la relación entre los picos de ventas con movimientos en la red social Twitter donde hemos clasificado y agrupado por palabras clave, los tuits...
- Describir las ventas por comunidades autónomas y marcas de forma cronológica desde… hasta…

-Observar el comportamiento de los usuarios que utilizan Valenbisi según las variables meteorológicas.

-Comprobar si el comportamiento de los usuarios que utilizan bicicletas varía entre barrios periféricos y barrios del centro, e incluso entre barrios con una densidad de población semejante.

-Analizar si existen grupos definidos en los datos de nuestro proyecto (*clustering*).

-Seleccionar una estación y una hora y ver la probabilidad de que haya un borne libre o una bicicleta sin ocupar (*predicción*).

-Analizar la demanda de bicicletas en las estaciones de Valenbisi teniendo en cuenta los lugares de interés cercanos a estas estaciones y valorando si cae en fines de semana o festivos.

3. Calidad y Análisis exploratorio.

Describe la calidad de los datos y los resultados del análisis exploratorio efectuado. Explica el trabajo técnico como, por ejemplo, estadísticas aplicadas, visualizaciones o representaciones que has utilizado (puedes poner alguna captura como ejemplo), etc. Valora este esfuerzo del análisis exploratorio de tu proyecto.

No se trata de describir los objetivos resultado del proyecto, sino lo que has tenido que hacer para entender los datos, ver qué nos ofrecen, cómo de "sucia" es la fuente, etc.

Puedes comentar el descarte de datos de las fuentes. Aquellos que no vayas a tener en cuenta por no ser útiles a tus objetivos.

Recomendamos utilizar RMarkDown ya que el análisis exploratorio se espera que lo hayáis hecho en R. Poned un anexo donde veamos las instrucciones, comandos, funciones, etc. además de las gráficas.

Calidad de los datos

La calidad de nuestros datos se puede describir refiriéndonos a la precisión, integridad, relevancia y consistencia de los mismos. Basándonos en esto, podemos decir que la precisión de nuestros datos es correcta y exacta, ya que están libres de errores y abarcan todo el abanico de posibilidades de análisis a los que podemos acceder con nuestro tema. Así pues, nuestros datos son íntegros puesto que están completos, aunque los hayamos alterado y manipulado para eliminar aquella información que no nos aportaba para realizar el proyecto, y la relevancia de estos es adecuada para el propósito para el que se están utilizando, ya que nos son útiles y aplicables para los análisis que queremos realizar. Por último, la consistencia de los datos es notable debido a que estos son uniformes y coherentes, al cumplir que no tienen contradicciones.

Resultados del análisis exploratorio

Al realizar el análisis exploratorio, hemos podido detectar que nuestra base de datos final no presenta valores faltantes en ninguna de las variables que disponemos.

Además, hemos encontrados incrementos y decrementos anómalos en los datos de una hora respecto a otra. Para ver si este comportamiento es sistemático, se analizará el comportamiento de un grupo concreto de estaciones a lo largo de, por ejemplo, 2 semanas

Por su parte, se han realizado una serie de análisis a nivel de barrios y distritos. Como consecuencia, se puede concluir que la Carrasca es el barrio con más estaciones (14) y, a nivel distrito, se concluye que Quatre Carreres es el que presenta más estaciones (27).

Algunos gráficos y los códigos de R utilizados tanto para llevar a cabo el análisis como para el tratamiento previo de los datos pueden ser observados en el link del anexo.

Valoración del esfuerzo del AED

Realizar el análisis exploratorio de los datos previo nos ha servido para entender mejor los datasets que empleamos en nuestro proyecto, por lo que podemos calificar como óptima la realización de dicho AED puesto que nos ha proporcionado conclusiones que pueden influenciar en gran medida en la continuación de nuestro trabajo.

Descarte de datos de las fuentes

Teniendo en cuenta que tenemos una gran cantidad de datos tanto en nuestra base de datos principal como en los datasets complementarios a esta, hemos tenido que dejar de lado variables que no nos aportaban información relevante para llegar a nuestros objetivos.

Comenzando por nuestra base de datos principal, hemos desechado columnas que no contenían datos al dataset *de Valenbisi*, como son "*created_user"*, "*created_date*", "*last_edited_user*" y "*last_edited_date*". Las variables previas no nos aportaban nada al no tener datos para complementar a las estaciones y, por ello, nos planteamos que era mejor descartarlas.

Por una parte, en el caso de la base de datos "aemet" hemos descartado las variables que trataban el indicativo de la estación, el nombre de la estación, la provincia y la altitud. Por otra parte, en el dataset "festivos" no hemos descartado nada, mientras que en el dataset "barrios", hemos decidido eliminar el punto medio del barrio como coordenada, y las columnas "*gis_gis_barrios_area*", "*linkid*" y "*globalid*".

Anexo de comandos y gráficas

https://upvedues-my.sharepoint.com/personal/jluqsai_upv_edu_es/Documents/2B/PROY%20II/PROYII/rMarkdown/tratamiento.html