Data Science Introduction Exercises (20 points)

Ask a same question both in ChatGPT and Gemini, then compare the results. -- 5pt

The question I picked was:

can you explain why the death of douglas adams was such a loss?

This is the response by ChatGPT 3.5 on April 13th, 2024:

> Douglas Adams' death was significant because he was not only a brilliant writer but also a unique and influential voice in science fiction and comedy. He is best known for his series "The Hitchhiker's Guide to the Galaxy," which started as a radio series in 1978 and later expanded into novels, a television series, stage shows, comic books, and a feature film.
>
> Adams had a rare talent for blending sharp wit, philosophical insights, and absurdity in his writing. His work was not just entertaining but also thought-provoking, often exploring complex ideas about the universe, technology, and the nature of existence in a humorous and accessible way. He had a knack for turning mundane situations into absurd adventures and finding humor in the absurdities of everyday life.
>
> Beyond his literary contributions, Adams was also an advocate for environmental causes and technology education. He was passionate about conservation and served as a dedicated spokesperson for organizations like Save the Rhino International.
>
> Adams' death in 2001 at the age of 49 was a loss not only for literature but also for the broader cultural landscape. His unique perspective and creative genius continue to inspire writers, comedians, and fans around the world.

Whereas here is the respons by Geminy, at the same date:

> Douglas Adams' death was a loss for several reasons:
>
> - His writing was immensely popular. The Hitchhiker's Guide to the Galaxy series is a beloved science fiction comedy that has sold millions of copies and garnered a devoted fanbase. Adams' wit, absurdist humor, and insightful observations on life, the universe, and everything resonated with readers around the world.
>
> - He was a creative force. Adams wasn't just funny, his work tackled big ideas and philosophical questions in a way that was both entertaining and thought-provoking. He brought a fresh perspective to science fiction and inspired countless readers and writers.
>
> - His influence extended beyond books. Adams was an advocate for environmental conservation and a champion of technological innovation. He also foresaw the potential of the internet early on, even creating a website where users could contribute their own entries, a kind of early crowdsourced encyclopedia.
>
> His death at a relatively young age felt like the loss of a potential for many more works and contributions. Fans felt it as the end of an era of brilliant storytelling.

The response by Gemini appears to be more structured, due to the bulletpoint form. ChatGPT injects more facts into the response, giving more specific context regarding the life of Douglas Adams. With 198 vs. 177 words, ChatGPT is more elaborate.

I also asked ChatGPT 4 the same question, with this answer:

Douglas Adams, best known as the author of "The Hitchhiker's Guide to the Galaxy" series, was a brilliant mind who brought a unique blend of humor, wit, and philosophical insight to his writing. His death in 2001 at the age of 49 was a significant loss for several reasons:

1. Creative Genius: Adams had a rare talent for blending science fiction with comedy in a way that was both thought-provoking and entertaining. His work resonated with millions of readers worldwide, and his writing style influenced a generation of authors.

2. Cultural Impact: "The Hitchhiker's Guide to the Galaxy" series became a cultural phenomenon, spanning books, radio dramas, TV shows, and even a feature film. The series introduced iconic characters like Arthur Dent, Ford Prefect, and Marvin the Paranoid Android, and popularized catchphrases like "Don't Panic" and "42." Adams' unique perspective on life, the universe, and everything captivated audiences and left a lasting impact on popular culture.

3. Literary Legacy: Adams' writing wasn't just entertaining; it was also deeply insightful. Through his humorous narratives, he explored complex themes such as the nature of existence, the meaning of life, and humanity's place in the universe. His ability to tackle profound ideas with wit and satire made his work both accessible and thought-provoking.

4. Unfinished Projects: At the time of his death, Adams was working on several projects, including a new installment in the "Hitchhiker's Guide" series and a screenplay for a feature film adaptation of his novel "Dirk Gently's Holistic Detective Agency." His untimely passing left many fans mourning not only the loss of a beloved author but also the potential for future works that would never come to fruition.

Overall, Douglas Adams' death was a loss not only for literature but for the world at large. His unique voice, boundless imagination, and keen insight into the human condition continue to inspire and delight readers to this day.

With even more details and in a longer form of 316 words, and better structurization due to "sub-headers" for the bullet points, it seems to be both the most complete and the most useful answer, since I as the asker of the question can pick the most relevant sections.

Neither of the LLMs seems to hallucinate in the responses, this probably is because the event is documented well and thus contained in the training data.

The google Gemini website is slightly more explicit in communicating that LLMs might hallucinate, though google claims to have checked the generated output using google searches for correctness. A more rigorous or less fact-based question might turn up greater differences.

You are a CEO from a business/startup with around 50 employees. You are considering to hire one or several data positions, what kind of data employee are you going to hire and why. Please describe the business, product/service first, and explain the reason of hiring data experts. (you can also not hire anyone) -- 10pt

I'm extrapolating from the job situation I'm in currently, which is being a developer in a small family-owned business with about 20 employees, building software products for two different sectors: maritime management software to support shipping companies in managing their fleet, and insurance broker-specific office management software targeted at small to mid-sized client companies.

We are on the market for almost 35 years now, and I'm with the company for ten. This year, we put lots of effort into being able to provide the vessel management software in other languages apart from English, so that additional markets, specifically French, Korean, and Japanese speaking ones can be explored. In addition to that, lots of our current customers employ Russian-speaking seamen, so that expanding the translations into Russian is desired.

In addition to that, we are expanding the financial management software from a locally-installed windows application to browser based apps, opening up both tablet and smartphone usage and allowing for easier deployment.

In the last years, alone, we hired 3 to 4 additional developers, so these innovations make it plausible that our company will grow to a size of about 50 employees in the near-ish future.

Expanding on the fact that our applications allow our customers to manage their data, we have additional areas I can see us needing additional data positions. One such area would be the aggregation of the existing data our customers use – extrapolating and comparing data that is common for different customers, allowing us to provide better services. A rather trivial example for this is basis data that is being manually entered by each (or multiple) of our customers, certification accordance documentation, for example. Extrapolating common patterns, e.g. from personnel travel logs, and providing better estimations for when to send crew where, is another example.

Similar approaches can be found in the financial application, too. Obviously, not only GDPR dictates that the usage and agglomeration of personal identifiable information must be handled with great care.

So one core requirement for additional hires, in addition to obvious knowledge in technologies like SQL is an awareness for these concerns: data protection, anonymization, pseudonymization.

Given that our userbase doesn't consist of the greater public, but employees of individual customer companies, and that their systems are being deployed either on-premises or managed using our cloud infrastructure, the amount of data we need to be able to handle is manageable, since that data is input by humans, doing their normal business cases, like reporting on executed maintenances or handling insurance incidents. We are talking of tens to hundreds of gigabytes per customer, where most of the volume of the data stems from binary files that are being imported into our system.

This doesn't necessarily qualify as Big Data, so we can save investing in that area – there's nothing in the foreseeable future that cannot be tackled by using single SQL server instances. Thus, we don't need specialized talent in technologies like e.g. Hadoop.

Our focus when expanding our employee base will be on developers, primarily, though – we have backlogs of projects that keeps us busy for the coming decades without any problem, and, as a family-owned company, we don't aspire to grow rapidly and vastly. The approach would be first to expand the developer-base for the existing projects, by 3 or 4 per software branch, and then, hiring additional teams for additional projects. Those often lie in the field of creating interfaces to external systems and connecting them with our databases.

Here, finding and keeping talent proved to be the most crucial point. We've been using apprenticeships, internships and student jobs, extensively, in the past to probe people, and will continue to do that.

Once those steps have been taking, we are going to expand into the data-oriented kind of meta-level operations I exampled earlier. We will focus on hiring a team of business analysts first, allowing them to explore the potential business cases deeper. Then, in conjunction with our existing team of developers, we will require a team of data scientists to actually implement them.

When I say team, I mean 3 to 4 people, each. We would never hire one specific individual only to do a certain and rely on them being available all the time.